MontCAS, Phase 2
Criterion-Referenced Test

2007-08
Technical Report

Montana Office of Public Instruction
PO Box 202501
Helena, Montana 59620-2501
www.opi.mt.gov

TABLE OF CONTENTS

SECTION I: ASSESSMENT DEVELOPMENT

Chapter 1. BACKGROUND AND OVERVIEW

1.1 Purpose of This Report

In the spring of 2008, Montana students in grades 3 through 8 and 10 participated in the MontCAS, Phase 2 Criterion Referenced Test (Montana CRT) in reading, mathematics, and science. The purpose of this assessment is to measure their achievement as articulated by the Montana Content Standards and Grade Level Expectations. The 2007-08 CRT was the fifth year of the operational program.

The purpose of this report is to describe several technical aspects of the Montana CRT in an effort to contribute to the accumulation of validity evidence to support Montana CRT score interpretations. Because it is the interpretations and uses of test scores that are evaluated for validity, in addition to the test, this report presents documentation to substantiate intended interpretations (American Educational Research Association (AERA), American Psychological Association & National Council on Measurement in Education, 1999). Subsequent chapters of this report discuss test development, test alignment, test administration, scoring, equating, item analyses, reliability, scaled scores, performance levels, and reporting. Each of these topics contributes important information towards establishing the validity of the assessment program. Note however that certain aspects of a comprehensive validity argument are not included in the report that could also be important to consider when drawing conclusions about validity (e.g., additional sources of validity evidence might speak to the extent to which scores from the Montana CRT assessments converge with other measures of the same or similar constructs and diverge from measures of different constructs; consequences that arise from scores at the student, school, district and state levels).

Historically, some parts of technical reports may have been used by educators and other stakeholders, but the intended audience was experts in psychometrics and educational research. This

edition of the Montana CRT technical report is an attempt to make the information more accessible to educators and other stakeholders, by providing richer descriptions of general categories of information. In making some of the information more accessible, we have purposefully preserved the depth of technical information historically provided. The reader will find that some of the discussion and tables continue to require a working knowledge of measurement concepts such as "reliability" and "validity" and statistical concepts such as "correlation" and "central tendency." To understand fully some of the presented data, the reader will have to be familiar with basic understanding of advanced topics in measurement and statistics.

1.2 Overview of the Assessment System

The Montana CRT was developed in accordance with the following federal laws: Title 1 of the Elementary and Secondary Education Act (ESEA) of 1994, P.L. 103-382 and the No Child Left Behind Act (NCLB) of 2001.

The Montana grade-content CRT tests are based on, and aligned to, Montana's Content Standards, Benchmarks and Grade Level Expectations in reading, mathematics, and science. The 2007-08 administration of the Montana CRT science test represents its first operational year. Detailed information about the design of the Montana CRT science test may be found in Chapter 6. The standard setting report for the science test is included as Appendix C.

Montana educators worked with OPI and its contractor, Measured Progress, to develop test items to assess how well students have met the Montana content Grade Level Expectations. In addition, an independent alignment study was performed by Northwest Regional Educational Laboratory (NWREL) for mathematics and reading in 2006 and science in 2007. NWREL's alignment studies may be found at http://www.opi.mt.gov/assessment/Phase2.html#Align.

Montana CRT scores are intended to be useful indicators of the extent to which students have mastered material outlined in the Montana Reading,Mathematics, and Science Content Standards,

Benchmarks, and Grade Level Expectations. For a particular student, his or her Montana CRT score should be used as part of a body of evidence regarding mastery and should not be used in isolation to make high stakes decisions. Montana CRT scores may be more reliable indicators of performance when aggregated to school, system, or state levels, particularly when monitored over the course of several years.

Table 1-1. 2007-08 Montana CRT: Timeline of Major Program Milestones

Milestone	*Year*	*Subjects*
Montana Content Standards adopted by Montana's Board of Education	1998	Reading and Mathematics
Item development and field test administration of the grades 3 through 8 and 10 CRT Montana-specific items	2003	Reading and Mathematics
First operational administration of the CRT in grades 4, 8 & 10	2004	Reading and Mathematics
Standard Setting for grades 4, 8 and 10	2004	Reading and Mathematics
Second operational administration of the CRT in grades 4, 8 & 10	2005	Reading and Mathematics
field test administration in grades 3, 5, 6 and 7	2005	Reading and Mathematics
Third operational administration of the CRT in grades 4, 8 & 10; First operational administration of the CRT in grades 3, 5 6 & 7	2006	Reading and Mathematics
Standard Setting for grades 3 through 8 and 10	2006	Reading and Mathematics
Item development and bias review by Montana educators to prepare for science field test in spring 2007	2006	Science
Fourth operational administration of the CRT in grades 4, 8 & 10; Second operational administration of the CRT in grades 3, 5 6 & 7	2007	Reading and Mathematics
field test administration in grades 4, 8 and 10	2007	Science
Fifth operational administration of the CRT in grades 4, 8 & 10	2008	Reading, Mathematics, and Science
Third operational administration of the CRT in grades 3, 5 6 & 7	2008	Reading and Mathematics
Standard Setting for grades 4, 8 & 10	2008	Science

1.3 Options for Participation

All Montana students enrolled in accredited schools are expected to participate in either the Montana CRT or the Montana CRT-Alternate. The vast majority of students will participate in the CRT, and most of them will participate under standard administration procedures. However, there is an array of accommodations which are available to any student, with or without disabilities, when such accommodations are necessary to allow the student to demonstrate his/her skills and competencies. For a list of standard accommodations please see Appendix H.

Standard accommodations are not considered to change the construct being measured and may be provided to students for any or all of the reading, math, or science portions of the assessment as necessary. Students' tests are scored the same way regardless of whether or not they took the test using standard accommodations.

In addition to standard accommodations, "non-standard accommodations" are available to a student. Non-standard accommodations on the Montana CRT may be provided in reading, math, or science as dictated by the student's IEP, 504, or LEP plan. Non-standard accommodations are considered to alter the construct being measured and they do affect the student's score on the CRT. When a non-standard accommodation is used, the student's score for that content area is reported as the lowest possible (i.e., a scaled score of 200 will fall into the Novice performance level).

For a very small percentage of students, participation in the statewide assessment program will be achieved by participating in the CRT-Alternate. Students with significant cognitive disabilities who are working toward alternate academic achievement standards, as documented in their IEP plans, are eligible to take the CRT-Alternate. Technical characteristics of the CRT-Alternate program are described in a companion technical report.

Chapter 2. OVERVIEW OF TEST DESIGN

2.1 2.1 Criterion-Referenced Test (CRT)

The Montana CRT test items are developed and customized specifically for use on the CRT Montana and they are directly linked to Montana's Content Standards, benchmarks and Grade Level Learning Expectations which can be view at http://www.opi.state.mt.us/Accred/cstandards.html. The content standards are the basis for the reporting categories developed for each content area and are used to help guide the development of test items. No other content or process is subject to statewide assessment. An item may address part, all, or several of the benchmarks within a standard.

2.2 Item Types

Montana's educators and students were familiar with the item types that were used in the assessment program. The types of items used and the functions of each are described below.

- **Multiple-choice (MC)** items were used, in part, to provide breadth of coverage of a content area. Because they require no more than a minute for most students to answer, these items make efficient use of limited testing time and allow coverage of a wide range of knowledge and skills.
- **Short-answer (SA)** items were used to assess students' skills and their abilities to work with brief, well-structured problems that had one or a very limited number of solutions (e.g., mathematical computations). Short-answer items require approximately two minutes for most students to answer. The advantage of this type of item is that it requires students to demonstrate knowledge and skills by generating, rather than merely selecting, an answer.
- **Constructed-response (CR)** items typically require students to use higher-order thinking skills—evaluation, analysis, summarization, and so on—in constructing a satisfactory response. Constructed-response items should take most students approximately five to ten

minutes to complete. It should be noted that the use of released Montana CRT items to prepare students to answer this kind of item is appropriate and encouraged.

2.3 Common-Matrix Design

The Montana CRTs are structured using both *common* items and *field test* items (the latter are matrix-sampled.) The common items are taken by all students at a given grade level. Student scores are based only on the common items. In addition, a larger pool of matrix-sampled items is divided among the eight forms of the test at each grade level. The field test items were transparent to test takers and had a negligible impact on testing time. Each student takes only one form of the test and so answers a fraction of the matrix-sampled items in the entire pool. This embedded field test design provides the sample size needed to produce reliable data (750-1500 students per item) on which to inform item selection for future tests

Chapter 3. TEST DEVELOPMENT PROCESS

3.1 Montana CRT Item Development

The items developed for the Montana CRT are consistent the Montana content standards and grade level learning expectations. Measured Progress curriculum and assessment specialists worked with Montana educators to verify the alignment of items to the appropriate Montana content standards. As an additional quality control check, Northwest Regional Educational Laboratory (NWREL) performed an independent alignment study to verify item alignment to Montana content standards for mathematics and reading in 2006 and science in 2007.

The development process followed by Measured Progress combined the expertise of the item development team and a panel of Montana educators to help ensure that items met the needs of the core MPSSIP program and the CRT program. All items used in the MPSSIP common portions of the Montana CRT program underwent review by a Montana content panel and a bias review panel. Annual MPSSIP item development is depicted in the following tables:

Table 3-1. 2007-08 Montana CRT: Annual MPSSIP Total Item Development—Grades 3–8 and 10

Grade	*Mathematics*	*Reading*	*Science*
3	78	160	
4	78	160	116
5	78	160	
6	78	160	
7	78	160	
8	78	160	116
10	78	160	116

Table 3-2. 2007-08 Montana CRT: Annual MPSSIP Reading Item Development—Grades 3–8 and 10

Passages	*Multiple Choice*	*Constructed Response*
2 long literary passages	40	4
2 long informational passages	40	4
4 short literary passages	40	0
4 short informational passages	40	0
12	**160**	**8**

Table 3-3. 2007-08 Montana CRT: Annual MPSSIP Mathematics Item Development—Grades 3–8 and10

Multiple Choice	*Short Answer*	*Constructed Response*
68	4	6

3.2 Item Development Process Overview

An overview of the item development process for the common and matrix items, including conducting the field tests, follows.

Table 3-4. 2007-08 Montana CRT: Item Development Process Overview

Development Step	*Step Details*
Select reading passages and conduct external review for bias and sensitivity issues (December, 2005)	Measured Progress Curriculum and Assessment Specialists located potential reading passages. Reading passages were reviewed for bias and sensitivity issues before the development of reading item sets.
Develop items (January through May 2006)	Measured Progress Curriculum and Assessment Specialists developed reading item sets and mathematics items.
Item review for bias and sensitivity issues and content appropriateness (May 2006)	Panels of Montana educators reviewed reading, mathematics, and science matrix field test items for bias and sensitivity issues.
Edit items (summer 2006)	Montana Educator's editorial comments were incorporated at this time
Matrix field test items (spring 2007)	Embedded field test (matrix) items were administered to a sample of students (maximum of 1,500 students per item/8 forms per grade and content).
Item Selection Meeting (July 2007)	Measured Progress test developers and Montana educators reviewed the results of the Spring 2007 matrix field test and selected common items for the Spring 2008 operational CRT forms.
Operational test items (March 2008)	Items are now part of the common item set and used to determine student scores

3.3 Internal Item and Content Review

The lead or peer Curriculum and Assessment Specialist within the content specialty reviewed each item for

- item "integrity" includes item content and structure, appropriateness to designated content area, item format and clarity.
- appropriateness and quality of reading selections and graphics, and appropriateness of scoring guide descriptions and distinctions (as correlated to the item and within the guide itself).
- Scorability including evaluating whether the scoring guide adequately addresses performance on the item, and parallel language between the item and scoring guide.
- fundamental issues including the following:
    - what is the item asking?
    - is the indicated key the only possible correct answer?
    - is the constructed-response item scorable as written (are the correct words used to elicit the response defined by the guide)?
    - is the item complete (i.e., with scoring guide, content codes, key, grade level, and contract identified)?
    - is the item appropriate for the designated grade level?

3.4 External Item, Content and Bias Reviews

All MPSSIP and Montana-augmented items underwent the following external reviews:

- In fall 2006, MPSSIP National Bias and Content Review Committees reviewed common and matrix passages and items used for the 2007-08 administration during two, two-day meetings, held in Salt Lake City, UT.
- In early December 2006, common item sets were reviewed by Measured Progress content specialists and Montana educators. Feedback from the Montana content and bias reviews were incorporated into the final editing processes.

3.5 Item Editing

Editors reviewed and edited the items to ensure uniform style (based on *The Chicago Report of Style, 15th Edition*) and adherence to sound testing principles. These principles included the stipulation that items

- were correct with regard to grammar, punctuation, usage, and spelling;
- were written in a clear, concise style;
- contained unambiguous explanations for students as to what was required to attain a maximum score;
- were written at a reading level that would allow the student to demonstrate his or her knowledge of the tested subject matter regardless of reading ability;
- had appropriate answer options or score-point descriptors; and
- were free of potentially insensitive content.

3.6 Operational Test Assembly

Test assembly is the sorting and laying out of item sets into test forms. In order to accommodate the embedded matrix field test design, eight forms of each test were administered in grades 3 through 8 and 10. Criteria considered during this process included the following.

- **Content coverage/match to test design**. The curriculum specialist completed an initial sorting of items into sets based on a balance of content categories across sessions and forms, as well as a match to the test design (e.g., number of multiple-choice, short-answer, and constructed-response items). See chapters 4-6 for specific content information.
- **Visual balance**. Item sets were reviewed to ensure that each reflected a similar length and "density" of selected items (e.g., length/complexity of reading selections or number of graphics).

- **Option balance**. Each item set was checked to verify that it contained a roughly equivalent number of key options (As, Bs, Cs, and Ds).
- **Name balance**. Item sets were reviewed to ensure that a diversity of names was used.
- **Bias**. Each item set was reviewed to ensure fairness and balance based on gender, ethnicity, religion, socioeconomic status, and other factors.
- **Page fit**. Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Facing-page issues**. For multiple items associated with a single stimulus (a graphic or a reading selection), consideration was given to whether those items needed to begin on a left- or right-hand page, as well as to the nature and the amount of material that needed to be placed on facing pages. These considerations served to minimize the amount of page flipping required of the students.
- **Relationships between forms**. Sets of common items were placed identically in each version of the forms. Although matrix-sampled item sets differed from form to form, they took up the same number of pages in each form so that sessions and content areas began on the same page in every form. Therefore, the number of pages needed for the longest form often determined the layout of each form.
- **Visual appeal**. The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of white space, the density of the text, and the number of graphics.

3.7 Editing Drafts of Operational Tests

Any changes made during the test construction had to be reviewed and approved by the Curriculum and Assessment Specialist. Once a form had been laid out in what was considered its final form, it was reread to identify any final considerations, including the following:

- **Editorial changes**. All text was scrutinized for editorial accuracy, including consistency of instructional language, grammar, spelling, punctuation, and layout. Measured Progress's publishing standards are based on *The Chicago Report of Style, 15th Edition*.
- **Keying items**. Items were reviewed for any information that might "key" or provide information that would help students answer another item. Decisions about moving keying items were based on the severity of the key-in and the placement of the items in relation to each other within the form.
- **Key patterns**. The final sequence of keys was reviewed to ensure that the order appeared random (i.e., no recognizable pattern and no more than three of the same key in a row).

3.8 Braille and Large-Print Translation

Form 1 for grades 3 through 8, and 10 tests was translated into Braille by National Braille Press, a subcontractor that specializes in test materials for blind and visually impaired students. In addition, *Form 1* for each grade was adapted into a large-print version.

Chapter 4. DESIGN OF THE READING ASSESSMENT

4.1 Reading Specifications

As indicated earlier, the test blueprint/specifications for reading were based on MPSSIP and Montana's reading content standards, which identify five Montana content standards that apply specifically to reading and reading comprehension. Those content standards follow:

- **Reading Standard 1:** Students construct meaning as they comprehend, interpret, and respond to what they read.
- **Reading Standard 2:** Students apply a range of skills and strategies to read.
- **Reading Standard 3:** Students set goals, monitor, and evaluate their reading progress. (This standard cannot be measured with a traditional paper/pencil test.)
- **Reading Standard 4:** Students select, read, and respond to print and non-print material for a variety of purposes.
- **Reading Standard 5:** Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences.

4.2 Reading Item Types

The Montana CRT reading assessments in reading include a mix of multiple-choice (MC) and constructed-response (CR) items. CR items require students to write an answer consisting of several phrases or short sentences. Each type of item is worth a specific number of points in the student's total reading score as shown in Table 4-1.

Table 4-1. 2007-08 Montana CRT: Reading Item Types and Point Values

Type of Item	*Possible Score Points*
Multiple-Choice (MC)	0 or 1
Constructed-Response (CR)	0, 1, 2, 3, or 4

4.3 Reading Test Design

Table 4-2 shows the number of MC and CR reading items by grade and test session.

Table 4-2. 2007-08 Montana CRT: Number of Common Reading Itemsby Grade and Test Session

				Total	
Grade	*Session 1*	*Session 2*	*Session 3*	*MC*	*CR*
3 -8	21 MC, 1 CR	10 MC	21 MC, 1 CR	52	2
10	21 MC, 1 CR	15 MC	21 MC, 1 CR	57	2

Table 4-3 shows the distribution of points across content standards.

Table 4-3. 2007-08 Montana CRT: Point Distribution Specifications/Blueprint for Reading Test by Standard and Grade

Total Number of Points on the Common (Scored) Test
52 MC items + 2 CR items = 60 points

Percent Point distribution by content standard*

Montana Content Standards	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Standard 1	34%	34%	34%	34%	34%	34%	25%
Standard 2	30%	30%	30%	30%	30%	30%	32%
Standard 4	18%	18%	18%	18%	18%	18%	22%
Standard 5	18%	18%	18%	18%	18%	18%	22%

*Because percents are rounded to the nearest whole number, not all sums add to 100%.
Note: Standard 3 cannot be measured with a traditional paper/pencil test.

Target point distribution by content standard (and acceptable range of points)

Montana Content Standards	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Standard 1	20 (18-22)	20 (18-22)	20 (18-22)	20 (18-22)	20 (18-22)	20 (18-22)	16 (14-18)
Standard 2	18 (16-20)	18 (16-20)	18 (16-20)	18 (16-20)	18 (16-20)	18 (16-20)	20 (18-22)
Standard 4	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	14 (12-16)
Standard 5	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	14 (12-16)

Four-point items: Each test contains two 4-point constructed-response items. In any given year, the two items will measure two different standards. From year to year, those standards may change.
One-point items: The number of one-point items per content standard will vary from year to year depending on which two standards are measured by the four-point items. (The number of total points per standard falls within the acceptable range from year to year.)

4.4 Reading Passage Types

Reading passages included both long and short texts selected from sources that students at each grade level would be likely to encounter in their classroom or in their independent reading. No passages were written specifically for the assessment, but instead were collected from published works. Each passage is classified as one of three types described below.

- **Literary passages** are represented by a variety of genres—modern narratives; diary entries; drama; poetry; biographies; essays; excerpts from novels; short stories; and traditional narratives, such as fables, myths, and folktales.
- **Content passages** are primarily informational and often deal with the areas of science and social studies. They are drawn from such sources as newspapers, magazines, and books.
- **Practical passages** are functional materials that instruct or advise the reader—for example, directions, reference tools, or reports.

The main difference in the passages used for grades 3–8, and 10 was their degree of difficulty. The primary formula used by Measured Progress to determine readability is the Flesch-Kincaid Grade Level formula to compare sentence length vs. syllables per word. This formula is modified from Flesch Reading Ease to reflect a U.S. grade-school reading level. All passages were selected to be appropriate for the intended audience; however, the ideas expressed became increasingly more complex from grades 3 through grade 10.

The items related to these passages required students to demonstrate their skills in both literal comprehension, where the answer is stated explicitly in the text, and inferential comprehension, where the answer is implied by the text and/or the text must be connected to relevant prior knowledge to determine an answer. In addition, some items focused on the reading comprehension skills reflected in content standards. Items of this type required students to use these skills and strategies to answer items—for example, how to identify the author's principal purpose, such as to

persuade, entertain, or inform—and to demonstrate their understanding of how words and images communicate to readers. Tables 4-4 and 4-5 depict passage distribution and length for Grades 3–8 and Grade 10.

Table 4-4. 2007-08 Montana CRT:
Point Distribution by Reading Passage Typefor All Grades 3–8 and 10

Passage Type	*Passage Content*	Percent of Test	Points Distribution
Literary	Stories, poetry, and other forms of literature	50 %	30 points
Informational	Content and practical passages	50 %	30 points
			60 points

Passage Length	*Passage Type*	*Percent of Test*	*Points Distribution*
Long	One literary or one informational per session	50 %	30 points
Short	At least one literary and informational per session	50 %	30 points
			60 points

Table 4-5. 2007-08 Montana CRT:
Approximate Reading Passage Lengths

Grade Level	*Long Passage (number of words)**	*Short Passage (maximum word length)**
Grade 3	350-800	350
Grade 4	400-850	400
Grade 5	450-850	450
Grade 6	450-900	450
Grade 7	450-950	450
Grade 8	500-1,000	500
Grade 10	550-1,200	550

While every attempt is made to adhere to recommended grade-level word counts for long and short passages, the final decision in the passage selection process is based on extensive reviews by content experts and bias panels, as well as a careful analysis of the sophistication of language, complexity of concepts, and readability of each passage. Table 4-5 shows the approximate length of the passages selected for the CRT.

Table 4-6. 2007-08 Montana CRT: Reading Test
Passage Design by Session for All Grades 3 through 8 and 10

Passages	*Number of Items*
Session 1: Common Augmented & Embedded Matrix (field test) Items	
Common short passage A	7 MC
Embedded short passage A	7 MC
Common Long Passage A	12 MC, 1 CR
Session 1 Total	**26 MC, 1 CR**
Session 2: Common Augmented & Embedded Matrix (field test) Items	
Common Short passage B	7 MC
Common Short passage C	7 MC
Embedded long passage	12 MC, 1 CR
Session 2 Total	**26 MC, 1 CR**
Session 3: Common Augmented & Embedded Matrix (field test) Items	
Embedded Short passage B	7 MC
Common short passage D	7 MC
Common Long Passage B	12 MC, 1 CR
Session 3 Total	**26 MC, 1 CR**
Common (Scored) Total	**52 MC, 2 CR**
Test Total	**78 MC, 3 CR**

Chapter 5. DESIGN OF THE MATHEMATICS ASSESSMENT

5.1 Mathematics Specifications

Mathematics specifications/blueprint is based on Montana's mathematics content standards:

- Mathematics Standard 1: Problem Solving
- Mathematics Standard 2: Numbers and Operations
- Mathematics Standard 3: Algebra
- Mathematics Standard 4: Geometry
- Mathematics Standard 5: Measurement
- Mathematics Standard 6: Data Analysis, Statistics, and Probability
- Mathematics Standard 7: Patterns, Relations, and Functions

5.2 Mathematics Item Types

The Montana CRT mathematics test includes MC, SA, and CR items. SA items require students to perform a computation or solve a simple problem. CR items are more complex, requiring 8–10 minutes of response time. Each type of item is worth a specific number of points in the student's total mathematics score, as shown below.

Table 5-1. 2007-08 Montana CRT: Mathematics Item Types and Point Values

Type of Item	*Possible Score Points*
Multiple-Choice	0 or 1
Short-Answer	0 or 1
Constructed-Response	0, 1, 2, 3, or 4

5.3 Mathematics Test Design

Table 5-2 summarizes the number and types of items that were used in constructing the common portions of the Montana CRT mathematics tests for 2007-08, and whether or not the use of

calculators was permitted. It should be noted that the Montana educators who helped develop the Montana CRT acknowledged the importance of mastering arithmetic algorithms. At the same time, they understood that the use of calculators is a necessary and important skill in society today. Calculators can save time and prevent error in the measurement of some higher-order thinking skills, allowing students to deal with more sophisticated and intricate problems. For these reasons, calculators were permitted on some parts of the Montana CRT mathematics test and prohibited on others. (Students were allowed to use any calculator with which they were familiar.)

Table 5-2. 2007-08 Montana CRT: Number of Common Mathematics Items by Grade and Test Session

	Grades 3–6		Grades 7, 8, and 10	
Session	Calculator	Number of Items	Calculator	Number of Items
1	Not Allowed	18 MC 2 SA 1 CR	Not Allowed	14 MC 3 SA 1 CR
2	Not Allowed	19 MC 1 SA	Allowed	21 MC
3	Allowed	18 MC 1 CR	Allowed	20 MC 1 CR

MC = multiple-choice items SA = short-answer items CR = constructed-response items

The mathematics test design consists of 55 multiple-choice items, three 1-point short-answer items, and two 4-point constructed-response items for 66 total points. Point distributions by content standard and grade are shown in the following table:

Table 5-3. 2007-08 Montana CRT: Point Distribution Specifications/Blueprint for Mathematics Test by Grade

Total Number of Points on the Common (Scored) Test
55 MC items + 2 CR items = 66 points

	Percent (and actual) point distribution by content standard*						
Content Standards	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Problem Solving+Number and Operations	34%(22)	34%(22)	32%(21)	32%(21)	27%(18)	27%(18)	20%(13)
Algebra	12%(8)	12%(8)	12%(8)	12%(8)	12%(8)	12%(8)	16%(11)
Geometry	15%(10)	15%(10)	16%(11)	16%(11)	18%(12)	18%(12)	20%(13)
Measurement	15%(10)	15%(10)	12%(8)	12%(8)	12%(8)	12(8)%	12%(8)
Data Analysis, Probability, and Statistics	12%(8)	12%(8)	15%(10)	15%(10)	18%(12)	18%(12)	20%(13)
Patterns, Relations, and Functions	12%(8)	12%(8)	12%(8)	12%(8)	12%(8)	12%(8)	12%(8)

*Because percents are rounded to the nearest whole number, not all sums add to 100%.

Note that two strands each year in a grade are measured by constructed-response items. Thus, the number of one-point items in a strand will vary depending on whether the strand contains a four-point item that year.

5.4 Mathematics Depth of Knowledge

Each item on the Montana CRT mathematics test is assigned a Depth of Knowledge (DOK) level according to the cognitive demand of the item. Depth of Knowledge is not synonymous with difficulty. The Depth of Knowledge level rates the complexity of the mental processing a student will use to solve a problem. A description of each of the four levels is shown below.

- Level 1 (Recall) This level requires the recall of a fact, definition, term, simple procedure, application of a formula, or performance of a straight algorithmic procedure. Items at this level may require students to demonstrate a rote response.
- Level 2 (Skill/Concept) This level requires mental processing beyond that of a habitual response. These items often require students to make some decisions as how to approach a problem.
- Level 3 (Strategic Thinking) This level requires students to develop a plan or sequence of steps. These items are more complex and abstract than the items at the previous two levels.

These items may also have more than one possible answer and require students to use evidence, make conjectures, or justify their answers.

- Level 4 (Extend Thinking) This level requires planning, investigation, and complex reasoning over an extend period of time. Students are required to make several connections within and across content areas. This level may require students to design and conduct experiments. Due to the nature of Level 4 no items on the CRT are rated as extend thinking.

It is important that the Montana CRT mathematics assessment measure a range of Depth of Knowledge. Table 5-2 shows the percent and point ranges of the three Depth of Knowledge levels used on the CRT mathematics assessment.

Table 5-4. 2007-08 Montana CRT: Point Distribution by Depth of Knowledge (DOK) Level for Mathematics Test

DOK Level	*Percent Range*	*Point Range*
1	20% to 30%	13 to 20 points
2	60% to 75%	39 to 50 points
3	5% to 10%	4 to 8 points

Chapter 6. DESIGN OF THE SCIENCE ASSESSMENT

6.1 Science Test Specifications

The science specifications/blueprint is based on Montana's science content standards:

- **Science Standard 1:** Scientific Investigations. Students, through the inquiry process, demonstrate the ability to design, conduct, evaluate, and communicate results and reasonable conclusions of scientific investigations.
- **Science Standard 2:** Physical Science. Students, through the inquiry process, demonstrate knowledge of properties, forms, changes, and interactions of physical and chemical systems.
- **Science Standard 3:** Life Science. Students, through the inquiry process, demonstrate knowledge of characteristics, structures and function of living things, the process and diversity of life, and how living organisms interact with each other and their environment.
- **Science Standard 4:** Earth/Space Science. Students, through the inquiry process, demonstrate knowledge of the composition, structures, processes and interactions of Earth's systems and other objects in space.
- **Science Standard 5:** Impact on Society. Students, through the inquiry process, understand how scientific knowledge and technological developments impact communities, cultures and societies.
- **Science Standard 6:** Historical Development – Students understand historical developments in science and technology.
- The Montana science e standards were developed for instruction purposes. Thus, Measured Progress item developers assumed, with the approval of Montana item review committees, that the content assessed was learned through the "inquiry process" and is thus assessed indirectly on the CRT.

6.2 Science Item Types

The Montana CRT science assessments include MC and CR items. MC items require students to select the correct response from four choices, each item taking one minute on average to answer. CR items are more involved and require 8–10 minutes of response time. Each type of item is worth a specific number of points in the student's total science score, as shown below.

Table 6-1. 2007-08 Montana CRT: Science Item Types And Point Values

Type of Item	*Possible Score Points*
Multiple-Choice (MC)	0 or 1
Constructed-Response (CR)	0, 1, 2, 3, or 4

6.3 Test Design

Table 6-2 summarizes the number and types of items on the common-item portion of the 2007-08 Montana CRT science tests (which are used to compute student scores). Additionally, each test form had matrixed field-test items (25 MC and 1 CR) which did not affect a student's score.

Table 6-2. 2007-08 Montana CRT: Number of Common Science Items By Test Sessions for All Grades 4, 8, and 10

Grades	*Session 1*	*Session 2*	*Session 3*	*TOTAL*
4, 8, and 10	18 MC, 1 CR	17 MC	18 MC, 1 CR	53 MC, 2 CR

Table 6-3 shows the distribution of points and item types across the content standards.

Table 6-3. 2007-08 Montana CRT: Point Distribution Specifications/ Blueprint For Science Test by Standard for All Grades 4, 8, and 10

Total Number of Points on the Common (Scored) Test
53 MC items + 2 CR items = 61 points

Percent Point distribution by content standard	
Montana Standards	Grades 4, 8, and 10
1. Scientific Investigations	23%
2. Physical Science	23%
3. Life Science	23%
4. Earth/Space Science	23%
5. Impact on Society 6. Historical Development	8%
Point distribution by content standard	
Montana Standards	Grades 4, 8, and 10
1. Scientific Investigations	14
2. Physical Science	14
3. Life Science	14
4. Earth/Space Science	14
5. Impact on Society 6. Historical Development	5
Item Type by content standard	
Montana Standards	Grades 4, 8, and 10
1. Scientific Investigations	10 or 14 MC; 1or 0 CR
2. Physical Science	10 or 14 MC; 1 or 0 CR
3. Life Science	10 or 14 MC; 1 or 0 CR
4. Earth/Space Science	10 or 14 MC; 1 or 0 CR
5. Impact on Society 6. Historical Development	1 or 5 MC; 1 or 0 CR

The science test design consists of 53 multiple-choice items 2 four-point constructed-response items for a total of 61 points. As with the mathematics test, each year two different standards are measured by constructed-response items and so the number of MC items in a strand is adjusted accordingly.

SECTION II: TEST ADMINISTRATION

Chapter 7. TEST ADMINISTRATION

7.1 Responsibility for Administration

As indicated in the *Test Coordinator's Manual*, principals and/or their designated School Test Coordinators were responsible for the proper administration of the CRT. This report was used to ensure the uniformity of administration procedures from school to school.

7.2 Procedures

School Test Coordinators were instructed to read the *Test Coordinator's Manual* prior to testing, and to be familiar with the instructions given in the *Test Administrator's Manual*. The *Test Coordinator's Manual* provided each school with checklists to help prepare for testing. The checklists outlined tasks to be performed before, during, and after test administration. Along with providing these checklists, the *Test Coordinator's Manual* outlined the nature of the testing material being sent to each school, how to inventory the material, how to track it during administration, and how to return the material once testing was complete. It also contained information about including or excluding students. The *Test Administrator's Manual* included checklists for the administrators to prepare themselves, their classrooms, and their students for the administration of the test. The *Test Administrator's Manual* contained sections that detailed the procedure to be followed for each test session, and it contained instructions on preparing the material prior to giving it to the School Test Coordinator for its return to Measured Progress.

7.3 Test Administrator Training

OPI hosted a test administration workshop at Helena, Montana on February 5–6, 2008. The workshop was well attended, but attendance by system and school test coordinators was not mandatory. OPI and Measured Progress staff hosted six sessions covering test accommodations,

student information system (AIM) updates, CRT materials and administration, CRT-Alternate materials and administration, online reporting, and test security. Each session was presented six times so that participants could be educated on all facets of test administration.

In addition to the workshop and the distribution of the *2008 Test Coordinator's Manuals* and *Test Administrator's Manuals*, OPI and Measured Progress produced and distributed one audio PowerPoint presentation, "Spring 2008: CRT and CRT-ALT Overview and Update of System and School Test Coordinators" to each system and school test coordinator. Training materials and the audio PowerPoint presentation were also posted on OPI's Web site. The training CD allowed system and school test coordinators who were unable to attend pre-administration workshops to be exposed to the training material and provided a useful training tool to both the system and school personnel

7.4 Participation Requirements

All students were expected to participate in the CRT; however, the scores of students in the following categories were excluded from the calculation of averages:

- Foreign exchange students
- Students not enrolled in an accredited Montana school (for example home-schooled student)
- Students enrolled in a private accredited school
- Students enrolled in a private non-accredited school
- Students enrolled in a private non-accredited Title 1 school
- Students enrolled part-time (less than 180 hours) taking a mathematics or reading course
- First year in US LEP students were required to participate in the mathematics assessment only.
- Students who took the CRT using a "non-standard" accommodation.

A summary of this information is shown in the table below which was published in the *Test Administrator's Manual* and *Test Coordinator's Manual*.

Table 7-1. 2007-08 Montana CRT: Summary of Eligibility for Exclusion from the CRT

Excluded from averages	*MUST Participate*	*MAY Participate*
Foreign exchange students	Yes	
Students not enrolled in an accredited Montana school		Yes
Students enrolled in a private accredited school	Yes	
Students enrolled in a private non-accredited school		Yes
Students enrolled in a private non-accredited Title I school		Yes
Students enrolled part-time (less than 180 hrs.) taking a mathematics or reading course		Yes
Reading: first year in US LEP students		Yes
Mathematics: First year in US LEP students	Yes	

Information about the exclusion was coded in by staff after testing was completed in the Student Response Booklet, if applicable. The *Test Coordinator's Manual* and *Test Administrator's Manual* provided detailed instructions for coding exclusions and accommodations. In addition, testing exclusions were discussed thoroughly in the pre-administration training audio CD. (See Appendix F—Reporting Decision Rules).

7.5 Test Scheduling

The Montana CRT tests were given during the spring of 2008: reading and mathematics tests to grades 3 through 8 and 10, science to grades 4, 8 and 10, during the four-week period, March 3–26, 2008. Schools were able to schedule testing sessions at any time during this period, provided they followed the sequence in the scheduling guidelines detailed in *Test Administrator's Manual.* Schools were asked to schedule makeup testing of students who were absent from initial test sessions during this testing window.

The Montana CRT is an un-timed assessment; however, guidelines or ranges were provided in the *2008 Test Coordinator's Manual* and *2008 Test Administrator's Manual* based on estimates of the time it would take an average student to respond to each type of item that made up the test:

- multiple-choice items – 1 minute per item
- short-answer items – 2 minutes per item
- constructed-response items – 10 minutes per item

The provided scheduling guidelines suggested scheduling 45–55 minutes per test session (50–60 minutes for grade 10 students.) There were 3 sessions per content area, and it was suggested that a break occur in between each session to prevent fatigue.

While the guidelines for scheduling were based on the assumption that most students would complete the test within the time estimated, each test administrator was asked to allow additional time for students who needed it. If additional classroom space was not available for students who required additional time to complete the tests, schools were encouraged to consider using another space for the purpose, such as the guidance office. If other areas were not available, it was recommended that each classroom used for test administration be scheduled for the maximum amount of time.

7.6 Help Desk

To address testing concerns, Measured Progress established a help desk dedicated to the Montana CRT. Help desk support is an essential element to the successful administration of large-scale assessments. It provides a centralized location where individuals in the field can call a toll-free number to request assistance, report problems they are experiencing, or ask specific questions.

The Measured Progress help desk provided support during all phases of the testing window. It was staffed at varying levels based on need and volume and was available from 8:00 A.M. to 4:00 P.M. MST during the testing window. At a minimum, the help desk consisted of a product support specialist who was responsible for receiving, responding to, and tracking calls and e-mails, and routing issues to the appropriate person(s) for resolution. In addition, communications requiring a

higher level of program support were routed to the program manager and/or program assistant. We received 224 calls to our service center for the following issues:

- Additional materials orders – 60 calls
- Materials inventory questions – 15 calls
- Student ID label questions – 17 calls
- Test security – 12 calls
- General testing questions – 16 calls
- UPS Pickup – 26 calls
- UPS Tracking – 5 calls.
- Other – 73 calls. Examples of "other" are; return boxes were thrown away, how do I return my materials? Packing questions, how to code accommodations properly, what to do with extra return materials?

When possible, all calls and e-mails received during business hours were responded to immediately with resolution or were updated within hours of receipt.

SECTION III: DEVELOPMENT AND REPORTING OF SCORES

Chapter 8. SCORING

Scoring of MC, SA, and CR is an important process of any large-scale assessment. The following paragraphs define the scoring processes used for the Montana CRT.

8.1 Scanning

Months prior to test administration and subsequent scanning activities, the Measured Progress scanning department met with the program management team to determine decision rules and required specifications for scanning and imaging. The information gathered at these meetings was then used to develop a customized scanning program for Montana.

For the Montana CRT program, Measured Progress used the NCS 5000i scanners, which employ rapid, highly accurate scanning and imaging technology. They feature real-time quality control checks, such as duplex read, the printing of a unique identifying number on each sheet of each booklet, and on-line editing capability,

At the conclusion of testing, Montana schools shipped all test materials back to Measured Progress. To expedite the scanning and scoring process, student response booklets were express-shipped separately from other test materials. 74,459 student response booklets were logged in; identified with appropriate scannable, preprinted school information sheets; examined for extraneous materials; counted and batched by school and grade; and moved into the scanning area.

At scanning, the booklet bindings were removed so that the individual pages could pass through the scanners one at a time. Once cut, the sheets were put back in their proper boxes and placed in storage until needed for the scanning/imaging process.

Customized scanning programs for all scannables were prepared to selectively read the student response booklets and to format the scanned information electronically according to predetermined requirements.

All student response documents and other scannable information necessary to produce the required reports were captured and converted into electronic format, including all student identification, demographics, and responses. The digital image clip information of SA and CR responses allowed Measured Progress to replicate student responses on readers' monitors just as they appeared on the originals. Data processing, scoring, benchmarking data analysis, and reporting were all accomplished electronically without further reference to the originals.

8.2 Scanning Quality Control

The scanning hardware is continually monitored for conditions that cause the machine to shut down if standards are not met. It displays an error message and prevents further scanning until the condition is corrected. Areas monitored include document page and integrity checks, user-designed on-line edits, and internal checks of electronic functions.

In an effort to protect data integrity, Measured Progress operators perform a diagnostic routine before every scanning shift begins. In the rare event that the routine detects a photocell that appears to be out of range, that machine is re-calibrated and tested again. If the read is still not up to standard, field service engineer is called in for assistance.

As a final safeguard, spot checks of scanned files, bubble by bubble and image by image, were routinely made throughout scanning runs. The result of all precautions was a scan error rate well below 1 per 1000.

8.3 Electronic Data Files

Test booklets were put into storage when scanning was complete; they are kept for at least 180 days beyond the close of the fiscal year. Once scanned files were determined to be complete and accurate, they were duplicated electronically and made available for many other processing options. Files were loaded onto the local area network (LAN) for transfer to Measured Progress's proprietary *iScore* system for scoring and used to identify and print papers to be used in the benchmarking processes, and the data were made transferable via the Internet, CD-ROM, or optical disk.

Table 8-1. 2007-08 Montana CRT: Number of Responses Scanned and Scored

Grade/Content	*Number of Responses Scanned and Scored*
3 Mathematics	89,647
4 Mathematics	81,043
5 Mathematics	79,375
6 Mathematics	78,457
7 Mathematics	81,936
8 Mathematics	85,478
10 Mathematics	89,165
3 Reading	35,006
4 Reading	34,712
5 Reading	32,040
6 Reading	35,412
7 Reading	35,528
8 Reading	37,037
10 Reading	39,234
4 Science	33,248
8 Science	35,004
10 Science	35,480

8.4 Items Scored by Readers

All Measured Progress scoring facilities use the Web-based, proprietary *iScore* process to score SA and CR items. *iScore* ensures the security of responses and test items: All scoring is "blind": No student names are associated with viewed responses or raw scores, and all scoring personnel are subject to the same nondisclosure requirements and supervision as regular Measured

Progress staff. Images of student responses are transferred electronically via a secure Web site to a scorer's computer screen at any one of Measured Progress's scoring facilities. For Montana's CRT program, scoring took place in Dover, New Hampshire, Albany, New York, Louisville, KY and Longmont, Colorado.

When *iScore* sends an image of a test response to an individual reader's computer terminal, the reader evaluates the response and records a score via keypad or mouse entry. A new response appears immediately on screen. The system guarantees complete anonymity of individual students and ensures the randomization of responses during scoring.

Although *iScore* is based on conventional scoring techniques, it also offers the following benefits;

- real-time information on scorer reliability, read-behinds, and overall process monitoring
- early access to subsets of data for tasks such as standard setting
- reduced material handling, which saves time and labor and enhances the security of materials
- immediate access to samples of student responses and scores for reporting and analysis through electronic media

Scoring operations were directed by the Montana CRT scoring project manager and carried out by the following staff:

- Chief Readers, who oversaw all training and scoring within particular subject areas
- Quality Assurance Coordinators (QACs), who led benchmarking and training activities and monitored scoring rates and consistency
- Senior Readers (SRs), who performed read-behinds of readers and assisted at scoring tables as necessary
- Readers, who performed the bulk of the scoring

Table 8-2 summarizes the educational credentials of the 2007-08 Montana CRT Readers and QACs.

Table 8-2. 2007-08 Montana CRT: Educational Credentials of Readers and QACs

Description	*Readers*				*Total*	*Pct*
	Albany, NY	*Denver, CO*	*Dover, NH*	*Louisville, KY*		
Less then 48 college credits	0	0	0	0	0	0.00%
48+ college credits	7	0	0	3	10	3.44%
Associate's degree	6	0	1	6	13	4.47%
Bachelor's degree	52	10	10	108	180	61.86%
Master's degree	27	1	5	38	71	24.40%
Doctorate	6	1	0	10	17	5.84%
Total	**98**	**12**	**16**	**165**	**291**	**100.01%**

Description	QACs				Total	Pct
	Albany, NY	Denver, CO	Dover, NH	Louisville, KY		
Less then 48 college credits	0	0	0	0	0	0.00%
48+ college credits	1	0	0	0	1	1.72%
Associate's degree	0	0	0	1	1	1.72%
Bachelor's degree	8	3	4	19	34	58.62%
Master's degree	4	2	3	11	20	34.48%
Doctorate	0	0	0	2	2	3.45%
Total	**13**	**5**	**7**	**33**	**58**	**99.99%**

8.5 Preliminary Activities

The preliminary activities for scoring included the following;

- participating in the planning and design of documents to be used for scoring
- reviewing items and score guides for benchmarking and training
- creating benchmarking packets
- selecting scoring staff and training them for scoring

8.6 Planning and Designing Documents

At the request of the scoring project manager, scoring personnel advised project management and OPI staff on the program design in order to support an efficient and effective scoring process. Scoring staff also contributed to the design of

- response documents, image-capturing process, file format and layout (in order to yield acceptable image clips);
- scoring benchmarks (a guide, subject background information, and anchor papers).

8.7 Benchmarking

Before the scheduled start of 2007-08 Montana CRT scoring activities, scoring center staff and Montana educators reviewed test items and scoring guides for benchmarking. At that point, Chief Readers and selected QACs prepared scorer training materials.

Scoring staff from Measured Progress, test developers, and Montana educators selected one or two *anchor* examples for each item score point. An additional six to ten responses per item were chosen as part of the *training* pack. The anchor pack consisted of midrange exemplars, while the training pack exemplars illustrated the full range within each score point. Chief Readers, who work closely with QACs for each content area, facilitated the selection of response exemplars.

8.8 Selecting and Training Scoring Staff

8.8.1 Quality Assurance Coordinators (QACs) and Senior Readers (SRs)

Because read-behinds by QACs and SRs moderate the scoring process and maintain the integrity of scores, the individuals chosen to fill these positions are selected for their accuracy. The QACs, who train readers to score each item in their content areas, are also selected for their ability to instruct and for their level of expertise in the content area. As such, QACs typically are retired

teachers who have demonstrated a high level of expertise in their disciplines. The ratio of QACs and SRs to readers was approximately 1:11.

8.8.2 Training QACs and SRs

To ensure that all QACs provided consistent training and feedback, Chief Readers spent two days training and qualifying the QACs, following which QACs reviewed all items with SRs. During scoring, QACs would rotate among tables, supervising Readers and reading behind SRs, who in turn read behind a different table of Readers each day.

8.8.3 Selecting Readers

Applicants for Reader were required to demonstrate their ability by participating in a preliminary scoring evaluation. The *iScore* system enables Measured Progress to measure efficiently a prospective Reader's ability to score student responses accurately. After participating in a training session, applicants were required to achieve at least 80% exact scoring agreement for a qualifying pack consisting of 20 responses to a predetermined item in their content area. The 20 responses were randomly selected from a bank of approximately 150 selected by QACs and approved by the CRs and item developers. Table 8-3 depicts the accuracy and qualification percentages of the Reader applicants.

Table 8-3. 2007-08 Montana CRT: Scoring Accuracy and Qualification Statistics

Grade-Content	*Item*	*Average % Exact Agreement for Embedded CR sets*	*Average % Exact Agreement for Double Blind Scoring*	*Number of Readers taking Qualification Sets*	*Number Successfully Qualifying*	*Percent Successfully Qualifying*
3 Mathematics	23	NA	83.3	NA	NA	NA
	24	NA	87.7	NA	NA	NA
	25	94.6	91.6	9	9	100.0
	48	NA	97.7	NA	NA	NA
	72	88.0	85.4	13	13	100.0
4 Mathematics	23	NA	97.1	NA	NA	NA
	24	NA	96.7	NA	NA	NA
	25	88.3	79.7	22	22	100.0
	48	NA	88.1	NA	NA	NA
	72	84.9	79.9	21	21	100.0

(cont'd)

Grade-Content	Item	Average % Exact Agreement for Embedded CR sets	Average % Exact Agreement for Double Blind Scoring	Number of Readers taking Qualification Sets	Number Successfully Qualifying	Percent Successfully Qualifying
5 Mathematics	23	NA	91.8	NA	NA	NA
	24	NA	96.8	NA	NA	NA
	25	NA	82.7	22	22	100.0
	48	NA	91.9	NA	NA	NA
	72	89.5	76.1	18	17	94.4
6 Mathematics	18	NA	89.4	NA	NA	NA
	19	NA	92.5	NA	NA	NA
	20	NA	93.3	NA	NA	NA
	23	84.6	86.9	11	9	81.8
	73	88.3	82.0	9	9	100.0
7 Mathematics	18	NA	93.1	NA	NA	NA
	19	NA	94.3	NA	NA	NA
	20	NA	97.8	NA	NA	NA
	23	89.2	93.4	28	26	92.9
	73	88.4	87.9	24	24	100.0
8 Mathematics	18	NA	90.8	NA	NA	NA
	19	NA	92.9	NA	NA	NA
	20	NA	92.7	NA	NA	NA
	23	89.9	90.5	20	19	95.0
	73	87.8	84.4	20	20	100.0
10 Mathematics	18	NA	92.0	NA	NA	NA
	19	NA	96.3	NA	NA	NA
	20	NA	92.5	NA	NA	NA
	23	88.3	92.3	22	22	100.0
	73	90.8	94.6	20	19	95.0
3 Reading	27	88.4	78.6	24	23	95.8
	81	79.9	71.1	24	24	100.0
4 Reading	27	87.5	78.9	23	21	91.3
	81	75.6	75.5	24	22	91.7
5 Reading	27	81.3	70.4	24	21	87.5
	81	87.1	73.8	23	23	100.0
6 Reading	27	85.6	71.7	22	19	86.4
	81	83.1	67.1	23	23	100.0
7 Reading	27	91.1	59.1	20	15	75.0
	81	86.3	66.1	17	16	94.1
8 Reading	27	90.2	61.8	21	19	90.5
	81	83.0	66.1	25	25	100.0
10 Reading	27	85.5	65.8	20	19	95.0
	81	91.2	71.0	30	30	100.0
4 Science	27	94.7	76.7	18	17	94.4
	81	88.9	78.8	19	19	100.0
8 Science	27	77.1	74.6	50	48	96.0
	81	76.0	69.5	28	27	96.4
10 Science	27	95.6	89.3	29	29	100.0
	81	64.2	91.2	31	31	100.0

8.8.4 Training of Readers

QACs commenced the actual training of Readers by demonstrating how to apply the language of the scoring guide for an item to its anchor pack exemplars. Following this, Readers scored the training pack. QACs reviewed the results of training pack scoring with Readers and answered their questions.

Tables 8-4 and 8-5 are examples of SA and CR scoring guides.

Table 8-4. 2007-08 Montana CRT: Short-Answer Item Scoring Guide

Score Point	*Description*
1	The student's response provides a complete and correct answer.
0	The student's response is totally incorrect or too minimal to evaluate.
B	Blank/no response.

Table 8-5. 2007-08 Montana CRT: Constructed- Response Item Scoring Guide

Score Point	*Description*
4	• The student completes all important components of the task and communicates ideas clearly. • The student demonstrates in-depth understanding of the relevant concepts and/or processes. • When instructed to do so, the student chooses more efficient and/or sophisticated processes. • When instructed to do so, the student offers insightful interpretations or extensions (e.g., generalizations, applications, and analogies).
3	• The student completes the most important components of the task and communicates clearly. • The student demonstrates understanding of major concepts even though he/she overlooks or misunderstands some less important ideas or details.
2	• The student completes most important components of the task and communicates those clearly. • The student demonstrates that there are gaps in his/her conceptual understanding.
1	• The student shows minimal understanding. • The student addresses only a small portion of the required task(s).
0	• The student's response is totally incorrect or irrelevant.
B	• Blank/no response.

Two aspects of scoring efficiency are in conflict with this system. First, in order to minimize training expense, it is desirable to train each Reader on as few items as possible. Second, to prevent reader drift and to minimize retraining requirements, it is desirable to score any given item within a

brief period of time. But the lower the number of unique items each Reader scores, the greater the number of Readers required to score that item quickly. To minimize this conflict, content-area Readers are divided into two or more groups. Groups are trained to score different items (or item sets). When they complete scoring all responses on that item, they are trained on another.

8.8.5 Monitoring Readers

Scoring of the 2007-08 Montana CRT took place over a period of approximately two weeks. Because items were randomly assigned to Readers, each item in an individual student's response booklet was more than likely scored by a different reader. This maximization of the number of Readers per each student booklet effectively minimizes the error variance due to Reader sampling.

As common and matrixed CR items were scored, two-percent of items were scored by SRs via "read-behind" at a rate of 2% of papers to ensure consistency across Readers and accuracy of individual Readers.

Individual Reader scores must exactly match the SR score more than 80% of the time and be at least adjacent 90% of the time. *iScore* is programmed to determine accuracy rates, and if a Reader is not meeting these standards, *iScore* alerts the SR. The SR determines whether that Reader's responses should be scored by another Reader, scored by a QAC, or routed for special attention. The SR also determines whether the Reader should continue scoring. Table 8-6 displays the final summary statistics for read-behind scoring, and Table 8-7 shows the actions taken with respect to Readers. SRs and QAC's were able to obtain current reader accuracy reports and speed reports online at any time.

Table 8-6. 2007-08 Montana CRT: Read-Behind Summary Statistics

Grade-Content	*Number of Responses Scored*	*Total Number of Responses Scored in Double-Blind*	*Total Number of Arbitrations Required*	*Percentage of Double-Blinds Arbitrated*
3 Mathematics	89,647	2,546	138	5.42%
4 Mathematics	81,043	2,660	94	3.53%
5 Mathematics	79,375	2,299	34	1.48%
6 Mathematics	78,457	3,056	80	2.62%
7 Mathematics	81,936	3,820	52	1.36%
8 Mathematics	85,478	3,946	87	2.20%
10 Mathematics	89,165	5,966	106	1.78%
3 Reading	35,006	1,059	18	1.70%
4 Reading	34,712	1023	23	2.25%
5 Reading	32,040	866	12	1.39%
6 Reading	35,412	946	24	2.54%
7 Reading	35,528	967	14	1.45%
8 Reading	37,037	1,019	18	1.77%
10 Reading	39,234	1,541	32	2.08%
4 Science	33,248	851	19	2.23%
8 Science	35,004	1,118	45	4.03%
10 Science	35,480	2,973	60	2.02%

To ensure high inter-rater reliability and to prevent scoring drift after a reader scored a student response, *iScore* determined whether the reader met the accuracy requirement which is that a reader's scoring, based on double-scored responses, must be exact more than 80% of the time and at least adjacent 90% of the time. If a reader's scores do not meet these standards, *iScore* will alert the senior reader, who will counsel the reader and determine if he/she should continue scoring. The senior reader will then determine whether responses should also be scored by another reader, scored by a QAC, or routed for special attention. QAC's and senior readers were able to obtain current reader accuracy reports and speed reports online at any time. Table 8-7 summarizes how often readers were prevented from scoring items through the qualification sets and quality control processes.

Table 8-7. 2007-08 Montana CRT: Blocked Reader Statistics

Grade-Content	*Grade/Item*	*Number of Readers Blocked From Scoring by* iScore	*Number of Readers NOT Allowed to Continue Scoring Based upon Other Quality Monitoring (Read-Behinds and Double Blinds)*
3 Mathematics	23	NA	NA
	24	NA	NA
	25	0	0
	48	NA	NA
	72	0	1
4 Mathematics	23	NA	NA
	24	NA	NA
	25	0	0
	48	NA	NA
	72	0	3
5 Mathematics	23	NA	NA
	24	NA	NA
	25	0	2
	48	NA	NA
	72	1	1
6 Mathematics	18	NA	NA
	19	NA	NA
	20	NA	NA
	23	1	1
	73	0	1
7 Mathematics	18	NA	NA
	19	NA	NA
	20	NA	NA
	23	2	2
	73	0	0
8 Mathematics	18	NA	NA
	19	NA	NA
	20	NA	NA
	23	1	0
	73	0	0
10 Mathematics	18	NA	NA
	19	NA	NA
	20	NA	NA
	23	0	2
	73	1	2
3 Reading	27	1	1
	81	0	1
4 Reading	27	2	2
	81	2	11
5 Reading	27	3	7
	81	0	3
6 Reading	27	3	2
	81	0	5

(cont'd)

Grade-Content	*Grade/Item*	*Number of Readers Blocked From Scoring by* iScore	*Number of Readers NOT Allowed to Continue Scoring Based upon Other Quality Monitoring (Read-Behinds and Double Blinds)*
7 Reading	27	5	2
	81	1	2
8 Reading	27	2	1
	81	0	7
10 Reading	27	1	3
	81	0	1
4 Science	27	1	0
	81	0	4
8 Science	27	2	25
	81	1	15
10 Science	27	0	0
	81	0	27

NOTE: All readers who were allowed to continue scoring did so under increased quality screening and additional read-behinds were conducted on these readers.

Chapter 9. ITEM ANALYSES

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each item, and both the *Standards for Educational and Psychological Testing (AERA et al., 1999)* and the *Code of Fair Testing Practices in Education (2004)* provide standards for identifying high-quality items. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors, for example. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Items must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative approaches are conducted to ensure that Montana CRT items meet these standards. Earlier sections of this report described the qualitative means by which item quality is assured. This section focuses on quantitative analyses, specifically, classical difficulty and discrimination indices, differential item functioning (DIF) statistics, dimensionality, and IRT statistics.

All analyses presented here are based on the statewide administration of the Montana CRT in spring 2008. The numbers of students who participated in the assessment at each grade level were about 10,300 in grade 3, 10,400 in grade 4, 10,300 in grade 5, 10,600 in grade 6, 10,600 in grade 7, 11,000 in grade 8, and 11,100 in grade 10.

The reader should keep in mind that the information presented in this chapter is based on *the items common to all forms* only, as it is these and only these items on which student scores are calculated. (Item analyses performed on the matrixed field-test items are used in the item review process and for purposes of form assembly in future administrations.)

9.1 Classical Difficulty and Discrimination Indices

All multiple-choice, constructed-response, and short-answer items were evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty was defined as the average proportion of points achieved on an item, and was measured by obtaining the average score on an item and dividing by the maximum possible score for the item. Multiple-choice items were scored dichotomously (correct vs. incorrect), so for those items, the difficulty index is simply the proportion of students who correctly answered the item. Constructed-response items (two on each mathematics form and two on each reading form) were scored polytomously, where a student can achieve a score of 0, 1, 2, 3, or 4. Short-answer items (three computation items on each mathematics form) were scored 0 or 1. By computing the difficulty index as the average proportion of points achieved, the indices for the different item types are placed on a similar scale; the index ranges from 0.0 to 1.0 regardless of the item type. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an *easiness* index, because larger values indicate easier items. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student ability, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students may indicate knowledge or skills that have not yet been mastered by most students, but such items provide little information about differences in student ability. In general, to provide best measurement, difficulty indices should range from near-chance performance (.25 for four-option, multiple-choice items or essentially zero for constructed-response or short-answer items) to .90. However, on a standards-referenced assessment such as the Montana CRT, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage (the Montana-CRT aims for a minimum of six items or points per standard).

Another desirable feature of an item is that the higher-achieving students perform better on the item than do lower-achieving students. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of an item. Within classical test theory, the item-test correlation is referred to as the item's discrimination, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For constructed-response items, the item discrimination index used was the Pearson product-moment correlation; for dichotomous items (multiple-choice and short-answer), this statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is –1.0 to +1.0 and their typical observed range is 0.2 to 0.6.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. Because each form of the Montana CRT was constructed to be parallel in content, the criterion score selected for each item was the raw score total for each form. The analyses were conducted for each form separately.

Summary statistics of the difficulty and discrimination indices for each item are provided in Tables 9-1 through 9-7 for grades 3 through 8 and 10. Mean difficulty and discrimination indices, broken down by item type—multiple-choice, constructed-response (which includes both the four-point constructed-response and one-point short-answer items), and all items—are shown in Table 9-8 (accompanied by standard deviations in parentheses). The item difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There were a small number of items with near-zero discrimination indices, but none were negative. Sometimes it

is necessary to include items with low discriminations or with very high or low difficulties to ensure that content is appropriately covered, but there were very few such cases on the Montana CRT.

A comparison of indices across grade levels is complicated because these indices are population dependent. Direct comparisons would require that either the items or students were common across groups. Since that is not the case, it can not be determined whether differences in performance across grade levels are due to differences in student ability or differences in item difficulty or both. However, one can say for mathematics that students in higher grades found their items more difficult than did students in lower grades.

Comparing the difficulty indices of multiple-choice items and constructed-response or short-answer items is inappropriate because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that the difficulty indices for multiple-choice items tend to be higher (indicating that students performed better on these items) than the difficulty indices for constructed-response items. Similarly, the partial credit allowed by four-point constructed-response items is advantageous in the computation of item-test correlations, so the discrimination indices for these items tend to be larger than the discrimination indices of multiple-choice or short-answer items.

Note that the descriptive statistics on difficulty and discrimination presented in Tables 9-1 through 9-7 and the summaries by item type in Table 9-8 are weighted according to the number of points contributed by each item.

Table 9-1. 2007-08 Montana CRT: Descriptive Statistics on Common-Item Difficulty and Discrimination Indices—Grade 3

Content Area		*Difficulty*	*Discrimination*
Reading	Mean	0.68	0.36
	StDev	0.16	0.08
	Min	0.27	0.14
	Max	0.94	0.53
	Range	0.67	0.39
Mathematics	Mean	0.70	0.37
	StDev	0.14	0.09
	Min	0.41	0.15
	Max	0.93	0.58
	Range	0.52	0.43

Table 9-2. 2007-08 Montana CRT: Descriptive Statistics on Common-Item Difficulty and Discrimination Indices—Grade 4

Content Area		*Difficulty*	*Discrimination*
Reading	Mean	0.65	0.37
	StDev	0.13	0.07
	Min	0.34	0.18
	Max	0.92	0.50
	Range	0.58	0.32
Mathematics	Mean	0.64	0.36
	StDev	0.16	0.09
	Min	0.27	0.20
	Max	0.90	0.57
	Range	0.63	0.37
Science	Mean	0.71	0.30
	StDev	0.17	0.07
	Min	0.27	0.12
	Max	0.94	0.49
	Range	0.67	0.37

Table 9-3. 2007-08 Montana CRT: Descriptive Statistics on Common-Item Difficulty and Discrimination Indices—Grade 5

Content Area		*Difficulty*	*Discrimination*
Reading	Mean	0.71	0.41
	StDev	0.15	0.07
	Min	0.33	0.20
	Max	0.93	0.55
	Range	0.60	0.35
Mathematics	Mean	0.61	0.36
	StDev	0.16	0.09
	Min	0.17	0.18
	Max	0.95	0.52
	Range	0.78	0.34

Table 9-4. 2007-08 Montana CRT: Descriptive Statistics on Common-Item Difficulty and Discrimination Indices—Grade 6

Content Area		*Difficulty*	*Discrimination*
Reading	Mean	0.70	0.37
	StDev	0.12	0.07
	Min	0.46	0.17
	Max	0.93	0.50
	Range	0.47	0.33
Mathematics	Mean	0.57	0.37
	StDev	0.17	0.09
	Min	0.22	0.19
	Max	0.92	0.65
	Range	0.70	0.46

Table 9-5. 2007-08 Montana CRT: Descriptive Statistics on Common-Item Difficulty and Discrimination Indices—Grade 7

Content Area		*Difficulty*	*Discrimination*
Reading	Mean	0.72	0.40
	StDev	0.11	0.07
	Min	0.45	0.22
	Max	0.89	0.52
	Range	0.44	0.30
Mathematics	Mean	0.53	0.35
	StDev	0.15	0.11
	Min	0.23	0.01
	Max	0.85	0.64
	Range	0.62	0.63

Table 9-6. 2007-08 Montana CRT: Descriptive Statistics on Common-Item Difficulty and Discrimination Indices—Grade 8

Content Area		*Difficulty*	*Discrimination*
Reading	Mean	0.73	0.38
	StDev	0.10	0.09
	Min	0.50	0.17
	Max	0.91	0.59
	Range	0.41	0.42
Mathematics	Mean	0.55	0.38
	StDev	0.14	0.10
	Min	0.29	0.20
	Max	0.84	0.67
	Range	0.55	0.47
Science	Mean	0.65	0.31
	StDev	0.16	0.09
	Min	0.29	0.04
	Max	0.92	0.53
	Range	0.63	0.49

Table 9-7. 2007-08 Montana CRT: Descriptive Statistics on Common-Item Difficulty and Discrimination Indices—Grade 10

Content Area		*Difficulty*	*Discrimination*
Reading	Mean	0.69	0.36
	StDev	0.12	0.08
	Min	0.49	0.16
	Max	0.95	0.57
	Range	0.46	0.41
Mathematics	Mean	0.48	0.35
	StDev	0.15	0.11
	Min	0.19	0.07
	Max	0.85	0.68
	Range	0.66	0.61
Science	Mean	0.57	0.34
	StDev	0.15	0.09
	Min	0.24	0.15
	Max	0.89	0.51
	Range	0.65	0.36

Table 9-8. 2007-08 Montana CRT: Means (Standard Deviations) of Common-Item Difficulty and Discrimination Indices and Number of Items, Overall and by Item Type for Each Grade-Content Combination

Grade	Content Area		Item Type		
			All	MC	CR
3	Reading	Difficulty	0.68 (0.16)	0.69 (0.16)	0.51 (0.02)
		Discrimination	0.36 (0.08)	0.36 (0.08)	0.47 (0.03)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.70 (0.14)	0.70 (0.14)	0.65 (0.11)
		Discrimination	0.37 (0.09)	0.36 (0.08)	0.39 (0.13)
		Number of Items	60	55	5
4	Reading	Difficulty	0.65 (0.13)	0.66 (0.13)	0.46 (0.02)
		Discrimination	0.37 (0.07)	0.37 (0.07)	0.41 (0.01)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.64 (0.16)	0.65 (0.16)	0.50 (0.12)
		Discrimination	0.36 (0.09)	0.34 (0.08)	0.48 (0.08)
		Number of Items	60	55	5
	Science	Difficulty	0.71 (0.17)	0.71 (0.17)	0.59 (0.16)
		Discrimination	0.30 (0.07)	0.29 (0.07)	0.42 (0.11)
		Number of Items	55	53	2
5	Reading	Difficulty	0.71 (0.15)	0.72 (0.14)	0.38 (0.02)
		Discrimination	0.41 (0.07)	0.40 (0.06)	0.48 (0.10)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.61 (0.16)	0.62 (0.16)	0.58 (0.18)
		Discrimination	0.36 (0.09)	0.35 (0.09)	0.39 (0.06)
		Number of Items	60	55	5
6	Reading	Difficulty	0.70 (0.12)	0.71 (0.11)	0.47 (0.01)
		Discrimination	0.37 (0.07)	0.37 (0.07)	0.46 (0.01)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.57 (0.17)	0.58 (0.17)	0.43 (0.18)
		Discrimination	0.37 (0.09)	0.36 (0.08)	0.48 (0.11)
		Number of Items	60	55	5
7	Reading	Difficulty	0.72 (0.11)	0.73 (0.1)	0.49 (0.04)
		Discrimination	0.40 (0.07)	0.40 (0.07)	0.49 (0.04)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.53 (0.15)	0.54 (0.15)	0.46 (0.16)
		Discrimination	0.35 (0.11)	0.33 (0.10)	0.51 (0.08)
		Number of Items	60	55	5
8	Reading	Difficulty	0.73 (0.10)	0.74 (0.10)	0.55 (0.04)
		Discrimination	0.38 (0.09)	0.38 (0.09)	0.53 (0.08)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.55 (0.14)	0.56 (0.15)	0.46 (0.07)
		Discrimination	0.38 (0.10)	0.36 (0.08)	0.57 (0.09)
		Number of Items	60	55	5
	Science	Difficulty	0.65 (0.16)	0.66 (0.16)	0.38 (0.08)
		Discrimination	0.31 (0.09)	0.31 (0.09)	0.48 (0.08)
		Number of Items	55	53	2
10	Reading	Difficulty	0.69 (0.12)	0.69 (0.12)	0.53 (0.06)
		Discrimination	0.36 (0.08)	0.35 (0.07)	0.55 (0.04)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.48 (0.15)	0.49 (0.15)	0.38 (0.13)
		Discrimination	0.35 (0.11)	0.33 (0.09)	0.55 (0.11)
		Number of Items	60	55	5
	Science	Difficulty	0.57 (0.15)	0.58 (0.15)	0.32 (0.07)
		Discrimination	0.34 (0.09)	0.33 (0.09)	0.50 (0.01)
		Number of Items	55	53	2

Note: Numbers shown in parentheses are standard deviations

9.2 Differential Item Functioning (DIF)

The *Code of Fair Testing Practices in Education* (2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit, and actions should be taken to make certain that differences in performance are due to construct-relevant, rather than irrelevant, factors. The *Standards for Educational and Psychological Testing* (AERA et al., 1999) includes similar guidelines. As part of the effort to identify such problems, Montana CRT items were evaluated in terms of differential item functioning (DIF) statistics.

DIF procedures are designed to identify items for which subgroups of interest perform differently beyond the impact of differences in overall achievement. For the Montana CRT, the standardization DIF procedure (Dorans and Kulick, 1986) was employed to evaluate subgroup differences for three comparison groups: male/female, white/Native American, and white/Hispanic. This procedure calculates the difference in item performance for groups of students matched for achievement on the total test. That is, the average item performance is calculated for students at every total score, then an overall average is calculated weighting by the total score distribution so the weighting is the same for the two groups. The index ranges from –1.00 to 1.00 for multiple-choice and short-answer items and is adjusted to the same scale for constructed-response items. Negative numbers indicate that the item was more difficult for female or non-white students. Dorans and Holland (1993) suggested that index values between –0.05 and 0.05 should be considered negligible. Most Montana CRT items fall within this range. Dorans and Holland further stated that items with values between –0.10 and –0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked, and that items with values outside the [–0.10, 0.10] range (i.e., "high" DIF) are more unusual and should be examined very carefully.

DIF indices indicate the degree of differential performance between two groups. That differential performance may or may not be indicative of bias in the test. Course-taking patterns, group differences in interests, or differences in school curricula can lead to DIF. If subgroup

differences in performance are related to construct-relevant factors, the items should be considered for inclusion on a test.

Each item was categorized according to the guidelines adapted from Dorans and Holland (1993). Table 9-9 shows the number of items classified into each category separately by item type (multiple-choice versus constructed-response; short-answer items are included with constructed-response). Results are shown for male/female, White/Native American, and White/Hispanic comparisons. Table 9-10 provides the number of items in each of the three DIF categories that advantaged males or females, also separately by item type (multiple-choice and constructed-response; constructed-response items are included with constructed-response). There are some Montana CRT items categorized as "low" or "high" DIF. These indices must not be interpreted as indisputable evidence of bias. Both the *Code of Fair Testing Practices in Education* (2004) and the *Standards for Educational and Psychological Testing* (AERA et al., 1999) assert that test items must be free from construct-irrelevant sources of differential difficulty. If subgroup differences in performance can be plausibly attributed to construct-relevant factors, the items may be included on a test. What is important is to determine if the cause of this differential performance is construct-relevant.

For the Montana CRT, there were relatively few items (less than five) flagged as having low or high DIF. The items that were flagged were reviewed for potential bias, and no obvious biases were detected. For this reason, and in order to ensure sufficient content coverage, no items were excluded from the test as a result of the DIF analyses.

Table 9-9. 2007-08 Montana CRT: DIF Analysis for Three Subgroup Comparisons, Overall and by Item Type, by Grade and Content Area

Grade	Content Area	*Male/Female DIF Class*									*White/Native American DIF Class*									*White/Hispanic DIF Class*								
		All			MC			CR			All			MC			CR			All			MC			CR		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
3	Reading	52	2	0	51	1	0	1	1	0	52	2	0	50	2	0	2	0	0	50	4	0	48	4	0	2	0	0
	Mathematics	56	4	0	51	4	0	5	0	0	57	3	0	52	3	0	5	0	0	54	6	0	49	6	0	5	0	0
4	Reading	51	3	0	49	3	0	2	0	0	46	8	0	44	8	0	2	0	0	52	2	0	50	2	0	2	0	0
	Mathematics	52	8	0	47	8	0	5	0	0	57	3	0	52	3	0	5	0	0	46	14	0	41	14	0	5	0	0
	Science	50	5	0	48	5	0	2	0	0	52	3	0	50	3	0	2	0	0	50	5	0	48	5	0	2	0	0
5	Reading	49	4	1	48	3	1	1	1	0	50	4	0	48	4	0	2	0	0	52	2	0	50	2	0	2	0	0
	Mathematics	45	15	0	43	12	0	2	3	0	58	2	0	53	2	0	5	0	0	55	5	0	50	5	0	5	0	0
6	Reading	49	5	0	48	4	0	1	1	0	47	6	1	45	6	1	2	0	0	49	5	0	47	5	0	2	0	0
	Mathematics	55	4	1	50	4	1	5	0	0	58	2	0	53	2	0	5	0	0	58	2	0	53	2	0	5	0	0
7	Reading	48	6	0	48	4	0	0	2	0	48	6	0	46	6	0	2	0	0	49	5	0	47	5	0	2	0	0
	Mathematics	50	9	1	46	8	1	4	1	0	57	3	0	53	2	0	4	1	0	48	10	2	43	10	2	5	0	0
8	Reading	46	8	0	45	7	0	1	1	0	53	1	0	51	1	0	2	0	0	48	6	0	46	6	0	2	0	0
	Mathematics	48	12	0	46	9	0	2	3	0	55	5	0	51	4	0	4	1	0	55	5	0	50	5	0	5	0	0
	Science	44	11	0	43	10	0	1	1	0	50	5	0	48	5	0	2	0	0	48	6	1	46	6	1	2	0	0
10	Reading	37	15	2	37	13	2	0	2	0	44	10	0	42	10	0	2	0	0	50	4	0	48	4	0	2	0	0
	Mathematics	50	10	0	47	8	0	3	2	0	56	4	0	52	3	0	4	1	0	51	9	0	46	9	0	5	0	0
	Science	47	8	0	46	7	0	1	1	0	51	4	0	49	4	0	2	0	0	50	5	0	48	5	0	2	0	0

A = negligible DIF, B = low DIF, C = high DIF

Table 9-10. 2007-08 Montana CRT: Male vs. Female DIF Categorization and Direction by Item Type, by Grade and Content Area

Grade	Content Area	Item Type	Negligible DIF (A)				Low DIF (B)				High DIF (C)			
			Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%
3	Reading	MC	28	23	51	98	0	1	1	2	0	0	0	0
		CR	1	0	1	50	1	0	1	50	0	0	0	0
	Mathematics	MC	35	16	51	93	1	3	4	7	0	0	0	0
		CR	4	1	5	100	0	0	0	0	0	0	0	0
4	Reading	MC	30	19	49	94	1	2	3	6	0	0	0	0
		CR	2	0	2	100	0	0	0	0	0	0	0	0
	Mathematics	MC	28	19	47	85	1	7	8	15	0	0	0	0
		CR	5	0	5	100	0	0	0	0	0	0	0	0
	Science	MC	24	24	48	91	1	4	5	9	0	0	0	0
		CR	2	0	2	100	0	0	0	0	0	0	0	0
5	Reading	MC	29	19	48	92	1	2	3	6	1	0	1	2
		CR	1	0	1	50	1	0	1	50	0	0	0	0
	Mathematics	MC	19	24	43	78	3	9	12	22	0	0	0	0
		CR	1	1	2	40	3	0	3	60	0	0	0	0
6	Reading	MC	26	22	48	92	2	2	4	8	0	0	0	0
		CR	1	0	1	50	1	0	1	50	0	0	0	0
	Mathematics	MC	29	21	50	91	1	3	4	7	0	1	1	2
		CR	5	0	5	100	0	0	0	0	0	0	0	0
7	Reading	MC	28	20	48	92	0	4	4	8	0	0	0	0
		CR	0	0	0	0	2	0	2	100	0	0	0	0
	Mathematics	MC	30	16	46	84	1	7	8	15	0	1	1	2
		CR	3	1	4	80	1	0	1	20	0	0	0	0
8	Reading	MC	27	18	45	87	1	6	7	13	0	0	0	0
		CR	1	0	1	50	1	0	1	50	0	0	0	0
	Mathematics	MC	25	21	46	84	4	5	9	16	0	0	0	0
		CR	1	1	2	40	2	1	3	60	0	0	0	0
	Science	MC	20	23	43	81	4	6	10	19	0	0	0	0
		CR	0	1	1	50	1	0	1	50	0	0	0	0
10	Reading	MC	23	14	37	71	4	9	13	25	1	1	2	4
		CR	0	0	0	0	2	0	2	100	0	0	0	0
	Mathematics	MC	25	22	47	85	3	5	8	15	0	0	0	0
		CR	3	0	3	60	2	0	2	40	0	0	0	0
	Science	MC	23	23	46	87	2	5	7	13	0	0	0	0
		CR	1	0	1	50	1	0	1	50	0	0	0	0

9.3 Dimensionality Analyses

The DIF analyses of the previous section were performed to identify items which showed evidence of differences in performance between pairs of subgroups beyond that which would be expected based on the primary construct that underlies total test score (also known as the "primary dimension;" for example, general achievement in mathematics). When items are flagged for DIF, statistical evidence points to their measuring an additional dimension(s) to the primary dimension.

Because tests are constructed with multiple content area subcategories, and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional IRT models that are used for calibrating, linking, scaling, and equating the 2007-08 Montana CRT test forms. As noted in the previous section, a statistically significant DIF result does not automatically imply that an item is measuring an *irrelevant* construct or dimension. An item could be flagged for DIF because it measures one of the construct-*relevant* dimensions of a subcategory's knowledge and skills.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Findings from dimensionality (DIM) analyses performed on the 2007-08 MontCAS common items for mathematics and reading are reported below. (Note: Only common items were analyzed since they are used for score reporting.)

Dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both of these methods use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on expected total score for the rest of the test, and the average conditional covariance is obtained by averaging over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected total test scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and

local *dependence* implies multidimensionality. Thus, non-random patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first divided into a training sample and a cross-validation sample. Then an exploratory analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items displays local dependence, conditioning on total score on the non-clustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first divided into a training sample and a cross-validation sample. (Note: The random training and cross-validation samples used for the DIMTEST analyses were drawn independently of the sample used for the DETECT analyses.) The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed, from this sum the between-cluster conditional covariances are subtracted, this difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality), values of 0.2 to 0.4 weak to moderate multidimensionality; values of 0.4 to 1.0 moderate to high multidimensionality, and values greater than 1.0 strong multidimensionality.

DIMTEST and DETECT were applied to the 2007-08 Montana CRT. The data for each grade and content area were split into a training sample and a cross-validation sample. Every grade-

content combination had at least 10,000 student examinees, so every training sample and cross-validation sample had at least 5,000 students. DIMTEST was applied to every grade-content. DETECT was applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

Because of the large sample sizes of the Montana tests, DIMTEST would be sensitive even to quite small violations of unidimensionality, and the null hypothesis was strongly rejected for every dataset ($p \leq 0.00005$ for every grade-content). These results were not surprising because strict unidimensionality is an idealization that almost never holds exactly for a given dataset. Thus, it was important to use DETECT to estimate the effect size of the violations of local independence found by DIMTEST. Table 9-11 displays the multidimensional effect size estimates from DETECT.

Table 9-11. 2007-08 Montana CRT: Multidimensionality Effect Sizes by Grade and Content Area.

Grade	*Content*	*Multidimensionality Effect Size*
3	Mathematics	0.13
	Reading	0.12
4	Mathematics	0.13
	Reading	0.11
	Science	0.09
5	Mathematics	0.18
	Reading	0.10
6	Mathematics	0.14
	Reading	0.13
7	Mathematics	0.12
	Reading	0.13
8	Mathematics	0.15
	Reading	0.12
	Science	0.08
10	Mathematics	0.13
	Reading	0.11
	Science	0.11

All the DETECT values indicated very weak multidimensionality. The mathematics and reading test forms (average effect size of about 0.13 and 0.12, respectively) tended to show slightly greater multidimensionality than did science (average of about 0.09). Also investigated was how DETECT divided the tests into clusters to see if there were discernable patterns with respect to item type (i.e., multiple choice and constructed response) or other factors. In grades 3, 4, 5, and 7 in mathematics, the constructed response items showed a slight to moderate tendency to cluster together, although the clusters also usually included multiple-choice items. No consistent clustering by item-type was found in any of the reading or science tests. The mathematics clusters showed no other discernable patterns. For all the reading grades and for grade 10 science, there was also some tendency for the items located near each other to cluster together. A more thorough type of investigation into identification of clusters that relate to the skills and knowledge areas measured by the items would need to employ experts in the substantive content of the test forms. In any case the violations of local independence from all such effects, as evidenced by the observed DETECT effect sizes on the 2007-08 Montana CRT, were very small and do not warrant any changes in test design or scoring.

9.4 Item Response Theory Analyses

In addition to the classical test theory item analyses reported earlier, the Montana CRT tests were analyzed according to item response theory (IRT) models. IRT analyses were used, first, to place all 2007-08 forms on the same scale, and second, to equate the 2007-08 test to the previous year's test. Details on the IRT calibration and equating procedures for the Montana CRT are provided in Chapter 11.

Chapter 10. RELIABILITY

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together and complement one another. Tests that function well provide a dependable assessment of the student's level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either higher or lower than his or her true ability. For example, a student may mis-read an item, or mistakenly fill in the wrong bubble when he or she knew the answer. Collectively, extraneous factors that impact a student's score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores or vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors made are small on average and student scores on such a test will consistently represent their ability) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable. (This is referred to as test-retest reliability.) A potential problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the "remembering items" problem is to give a different, but parallel test at the second administration. If student scores on each test correlate highly the test is considered reliable. (This is known as alternate forms reliability,

because an alternate form of the test is used in each administration.) This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel forms reliability indices. One way to address these problems is to split the test in half and then correlate students' scores on the two half-tests; this in effect treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval, and of creating and administering two parallel forms of the test, are alleviated. This is known as a split-half estimate of reliability. If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test halves will result in a different correlation. Another problem with the split-half method of calculating reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, alpha (α), which avoids these concerns of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2007–08 Montana CRT:

$$\alpha \equiv \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{n}\sigma^{2}_{(Y_i)}}{\sigma^{2}_{x}}\right]$$

Where
i indexes the item
n is the total number of items,
$\sigma^{2}_{(Y_i)}$ represents individual item variance, and
σ^{2}_{x} represents the total test variance

Another approach to estimating the reliability for a test with differing item types (i.e., multiple-choice and constructed-response) is to assume that at least a small, but important, degree of unique variance is associated with item type (Feldt and Brennan, 1989), in contrast to Cronbach's α, which assumes that there are no such local or clustered dependencies. A stratified version of coefficient α corrects for this problem by using the following formula:

$$\alpha_{strat} = 1-\frac{\sum_{j=1}^{k}\sigma^{2}_{x_j}(1-\alpha_j)}{\sigma^{2}_{x}}$$

where j indexes the subtests or categories,
$\sigma^{2}_{x_j}$ represents the variance of each of the k individual subtests or categories,
α_j is the unstratified Cronbach's α coefficient for each subtest, and
σ^{2}_{x} represents the total test variance.

10.1 Reliability and Standard Errors of Measurement

Table 10-1 provides descriptive statistics for the Montana CRT, Cronbach's α coefficient, and raw score standard errors of measurement for each grade and content combination. Table 10-2 presents Cronbach's α for each item type and then stratified on item type for each grade-content. Across the grades and content areas, the overall α coefficients, multiple-choice α coefficients, and stratified α coefficients range from the mid-.80s to the low-.90s. There is little or no difference between the overall α and stratified α coefficients. The α coefficients for the constructed-response

items are substantially lower, ranging from around 0.40 to around 0.70. These lower values can be explained, at least to some extent, by the fact that there are greater scoring inconsistencies for constructed-response items, as well as the relatively small numbers of these items on the test. Note that, for reading, it is possible that the reliability coefficients are inflated as a result of passage-based item dependency.

Table 10-1. 2007-08 Montana CRT: Common-Item Descriptive Statistics, α Reliability, and Standard Errors of Measurement by Grade and Content

Grade	*Content Area*	*N*	*Total Points*	*Mean*	*SD*	*Reliability*	*SEM*
3	Mathematics	10333	66	45.472	11.582	0.906	3.548
	Reading	10317	60	39.958	9.836	0.895	3.185
4	Mathematics	10356	66	41.272	11.531	0.904	3.568
	Reading	10330	60	37.764	10.314	0.901	3.248
	Science	10354	61	42.446	8.612	0.850	3.334
5	Mathematics	10305	66	39.858	11.696	0.899	3.709
	Reading	10280	60	40.651	10.488	0.916	3.033
6	Mathematics	10597	66	36.576	12.279	0.907	3.737
	Reading	10608	60	40.442	10.113	0.902	3.168
7	Mathematics	10646	66	34.185	12.114	0.899	3.848
	Reading	10652	60	41.766	10.718	0.917	3.086
8	Mathematics	10942	66	35.714	13.002	0.913	3.827
	Reading	10970	60	42.849	10.137	0.907	3.098
	Science	10957	61	37.888	9.484	0.869	3.434
10	Mathematics	11084	66	30.404	12.191	0.901	3.839
	Reading	11075	60	40.310	10.076	0.897	3.237
	Science	11072	61	33.275	10.569	0.887	3.560

Table 10-2. 2007-08 Montana CRT: Common-Item α Reliability Overall, by Item Type, and Stratified α, by Grade and Content

Grade	*Content Area*	*Overall α*	*MC α*	*Number of MC Items*	*CR α*	*Number of CR Items (tot CR pts)*	*Stratified α by Item Type*
3	Mathematics	0.91	0.90	55	0.47	5 (11)	0.90
	Reading	0.90	0.89	52	0.55	2 (8)	0.90
4	Mathematics	0.90	0.89	55	0.60	5 (11)	0.90
	Reading	0.90	0.90	52	0.52	2 (8)	0.91
	Science	0.85	0.85	53	0.38	2 (8)	0.86
5	Mathematics	0.90	0.90	55	0.46	5 (11)	0.90
	Reading	0.92	0.91	52	0.52	2 (8)	0.92
6	Mathematics	0.91	0.90	55	0.57	5 (11)	0.91
	Reading	0.90	0.90	52	0.64	2 (8)	0.91
7	Mathematics	0.90	0.88	55	0.61	5 (11)	0.90
	Reading	0.92	0.92	52	0.64	2 (8)	0.93
8	Mathematics	0.91	0.90	55	0.67	5 (11)	0.92
	Reading	0.91	0.90	52	0.69	2 (8)	0.91
	Science	0.87	0.86	53	0.45	2 (8)	0.87
10	Mathematics	0.90	0.88	55	0.66	5 (11)	0.90
	Reading	0.90	0.89	52	0.71	2 (8)	0.90
	Science	0.89	0.88	53	0.50	2 (8)	0.89

10.2 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2007-08 Montana CRT assessments. Appendix G presents reliabilities for various subgroups of interest. Subgroup Cronbach's α's were calculated using the formula defined above based only the members of the subgroup in question in the computations. For reading, subgroup reliabilities ranged from 0.79 to 0.92, for mathematics from 0.69 to 0.92, and for science from 0.75 to 0.92.

For several reasons, the results of this subsection should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but on the statistical distribution of the studied subgroup. For example, subgroup sample sizes may vary considerably (see Appendix G), resulting in natural variation in reliability coefficients. Alpha, like any other correlation coefficient, may be

artificially depressed for subgroups with little variability (Draper & Smith, 1998). Finally, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup.

10.3 Reporting Subcategories Reliability

In previous sections, the reliability coefficients were calculated based on form and item type. Item type represents just one way of breaking an overall test into subtests. Of even more interest are reliabilities for the reporting subcategories within Montana CRT subject areas, described in Chapters 4 through 6. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Table 10-3. Once again as expected, because they are based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this into account.

Table 10-3. 2007-08 Montana CRT:
Common Item α by Reporting Subcategory

Grade	*Subject*	*Reporting Subcategory*	*Possible Points*	α
3	Mathematics	Problem Solving + Numbers and Operations	22	0.81
		Algebra	8	0.51
		Geometry	10	0.50
		Measurement	10	0.53
		Data Analysis, Statistics, and Probability	8	0.52
		Patterns, Relations, and Functions	8	0.67
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	20	0.77
		Students apply a range of skills and strategies to read	21	0.75
		Students select, read and respond to print and non-print material for a variety of purposes	11	0.58
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	8	0.53
4	Mathematics	Problem Solving + Numbers and Operations	22	0.80
		Algebra	8	0.56
		Geometry	10	0.55
		Measurement	10	0.63
		Data Analysis, Statistics, and Probability	8	0.52
		Patterns, Relations, and Functions	8	0.44
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	21	0.79
		Students apply a range of skills and strategies to read	19	0.72
		Students select, read and respond to print and non-print material for a variety of purposes	10	0.55
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	10	0.64
	Science	Scientific Investigations	14	0.60
		Physical Science	14	0.58
		Life Science	14	0.54
		Earth/Space Science	14	0.57
		Impact on Society	2	0.13
		Historical Development	3	0.24
5	Mathematics	Problem Solving + Numbers and Operations	21	0.80
		Algebra	8	0.64
		Geometry	11	0.55
		Measurement	8	0.54
		Data Analysis, Statistics, and Probability	10	0.51
		Patterns, Relations, and Functions	8	0.40
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	24	0.79
		Students apply a range of skills and strategies to read	18	0.80
		Students select, read and respond to print and non-print material for a variety of purposes	9	0.60
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	9	0.65

Table 10-3. 2006-07 Montana CRT Common Item α by Grade, Subject, and Reporting Subcategory (cont'd).

Grade	*Subject*	*Reporting Subcategory*	*Possible Points*	α
6	Mathematics	Problem Solving +Numbers and Operations	21	0.78
		Algebra	8	0.65
		Geometry	11	0.53
		Measurement	8	0.49
		Data Analysis, Statistics, and Probability	10	0.62
		Patterns, Relations, and Functions	8	0.60
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	23	0.75
		Students apply a range of skills and strategies to read	22	0.81
		Students select, read and respond to print and non-print material for a variety of purposes	8	0.58
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	7	0.49
7	Mathematics	Problem Solving + Numbers and Operations	18	0.75
		Algebra	8	0.53
		Geometry	12	0.59
		Measurement	8	0.54
		Data Analysis, Statistics, and Probability	12	0.57
		Patterns, Relations, and Functions	8	0.58
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	21	0.78
		Students apply a range of skills and strategies to read	20	0.82
		Students select, read and respond to print and non-print material for a variety of purposes	9	0.52
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	10	0.64
8	Mathematics	Problem Solving + Numbers and Operations	18	0.75
		Algebra	8	0.70
		Geometry	12	0.57
		Measurement	8	0.58
		Data Analysis, Statistics, and Probability	12	0.65
		Patterns, Relations, and Functions	8	0.56
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	18	0.77
		Students apply a range of skills and strategies to read	21	0.75
		Students select, read and respond to print and non-print material for a variety of purposes	8	0.55
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	13	0.70
	Science	Scientific Investigations	14	0.58
		Physical Science	14	0.63
		Life Science	14	0.55
		Earth/Space Science	14	0.60
		Impact on Society	3	0.30
		Historical Development	2	0.31

Table 10-3. 2006-07 Montana CRT Common Item α by Grade, Subject, and Reporting Subcategory, (cont'd)

Grade	*Subject*	*Reporting Subcategory*	*Possible Points*	α
10	Mathematics	Problem Solving + Numbers and Operations	13	0.68
		Algebra	11	0.66
		Geometry	13	0.63
		Measurement	8	0.46
		Data Analysis, Statistics, and Probability	13	0.62
		Patterns, Relations, and Functions	8	0.51
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	17	0.71
		Students apply a range of skills and strategies to read	21	0.77
		Students select, read and respond to print and non-print material for a variety of purposes	11	0.57
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	11	0.63
	Science	Scientific Investigations	14	0.72
		Physical Science	14	0.64
		Life Science	14	0.59
		Earth/Space Science	14	0.59
		Impact on Society	3	0.35
		Historical Development	2	0.25

For reading, subcategory reliabilities ranged from 0.49 to 0.82, for mathematics from 0.40 to 0.81, and for science from 0.13 to 0.72. The subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect based on classical test theory. Qualitative differences between grades and content areas once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

10.4 Reliability of Performance Level Categorization

All test scores contain measurement error; thus classifications based on test scores are also subject to measurement error. After the performance levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the Montana CRT, students are classified into one of four performance levels: *Novice* (N), *Nearing Proficiency* (NP), *Proficient* (P), or *Advanced* (A). This

section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist.

Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete, parallel forms of the test are given to the same group of students. This is usually impractical, especially on lengthy tests. To overcome this issue, techniques have been developed to estimate both accuracy and consistency of classification decisions based on a single administration of a test. The technique developed by Livingston and Lewis (1995) was used for the Montana CRT because their technique can be used with both constructed-response and multiple-choice items.

All of the accuracy and consistency estimation techniques described below make use of the concept of "true scores" in the sense of classical test theory. A true score is the score that would be obtained on a test that had no measurement error. It is a theoretical concept that cannot be observed, although it can be estimated. In the Livingston and Lewis method, the estimated true score distribution is used to estimate the proportion of students in each "true" performance level. After various technical adjustments (which are described in Livingston and Lewis, 1995), a 4 × 4 contingency table was created for each content area test and grade level. The [*i*,*j*] entry of an accuracy table represents the estimated proportion of students whose true score fell into performance level *i* and whose observed score fell into performance level *j* on the Montana CRT. Overall accuracy, which is the proportion of students whose true and observed performance levels match one another, is the sum of the numbers on the diagonal of the accuracy table.

To estimate consistency, the true scores are used to estimate the joint distribution of classifications on two independent, parallel test forms. After statistical adjustments (see Livingston and Lewis, 1995), a new 4 × 4 contingency table was created for each test and grade level that shows the proportion of students who would be classified into each performance level by the two (hypothetical) parallel test forms. That is, the [*i*,*j*] entry of a consistency table represents the estimated proportion of students whose observed score on the first form would fall into performance level *i* and whose observed score on the second form would fall into performance level *j*. Overall consistency, which is the proportion of students classified into exactly the same performance level by the two forms of the test, is the sum of the numbers on the diagonal of this new contingency table.

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. Cohen's κ can be used to evaluate the classification consistency of a test from two parallel forms of the test. The two forms in this case were the hypothetical parallel forms used by the Livingston and Lewis method. Because κ is corrected for chance, the values of κ are lower than other consistency estimates.

10.5 Results of Accuracy, Consistency, and Kappa Analyses

Summaries of the Accuracy and Consistency analyses are provided in Tables 10-4 through 10-20. The first section of each table shows the overall accuracy and consistency indices as well as Kappa. The overall index is, as described above, the sum of the diagonal elements of the appropriate contingency table.

The second section of each table shows accuracy and consistency values conditional upon performance level. In each case, the denominator is the number of students who are associated with a given performance level. For example, the conditional accuracy value is 0.7770 for the *Proficient* category for Grade 4 mathematics. This indicates that, of the students whose true scores placed them

in the *Proficient* category, 77.770% of them would be expected to be in the *Proficient* category if they were categorized according to their observed scores. The corresponding consistency value of .7113 indicates that 71.13% of students with observed scores in the *Proficient* performance level would be expected to score in *Proficient* again if a second, parallel test form were used.

For certain tests, concern may be greatest regarding decisions made about a particular threshold. For example, if a college gave credit to students who achieved an Advanced Placement test score of four or five, but not one, two, or three, one might be interested in the accuracy of the dichotomous decision, below four versus four or above. The third section of the summary tables shows information at each of the cut points. These values indicate the accuracy and consistency of the dichotomous decisions, either above or below the associated cut point. In addition, the false positive and false negative accuracy rates are also provided. These values are estimates of the proportion of students who were categorized above the cut when their true score would place them below the cut (false positive), and vice versa.

Table 10-4. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 3 Reading

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.8445	0.7831	0.6347		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.7377	0.5516		
	Nearing Proficiency	0.7120	0.5920		
	Proficient	0.8273	0.7858		
	Advanced	0.9022	0.8329		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9905	0.0033	0.0062	0.9862
	NP : P	0.9535	0.0210	0.0255	0.9345
	P : A	0.9005	0.0614	0.0382	0.8618

Table 10-5. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 4 Reading

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.8113	0.7384	0.5993		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.7513	0.6198		
	Nearing Proficiency	0.7100	0.6111		
	Proficient	0.8112	0.7573		
	Advanced	0.8788	0.7953		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9728	0.0115	0.0157	0.9615
	NP : P	0.9335	0.0340	0.0324	0.9071
	P : A	0.9049	0.0588	0.0362	0.8683

Table 10-6. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 5 Reading

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.8180	0.7489	0.6147		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.7826	0.6661		
	Nearing Proficiency	0.6742	0.5649		
	Proficient	0.7782	0.7159		
	Advanced	0.9133	0.8489		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9739	0.0113	0.0149	0.9630
	NP : P	0.9371	0.0319	0.0310	0.9120
	P : A	0.9068	0.0583	0.0349	0.8712

Table 10-7. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 6 Reading

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.8108	0.7392	0.5991		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.8073	0.7121		
	Nearing Proficiency	0.6414	0.5250		
	Proficient	0.7731	0.7027		
	Advanced	0.8966	0.8341		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9716	0.0129	0.0155	0.9599
	NP : P	0.9427	0.0288	0.0285	0.9197
	P : A	0.8961	0.0608	0.0431	0.8558

Table 10-8. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 7 Reading

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.834	0.7702	0.6336		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.7971	0.6962		
	Nearing Proficiency	0.6619	0.5481		
	Proficient	0.7948	0.7311		
	Advanced	0.9122	0.8591		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9771	0.0104	0.0126	0.9676
	NP : P	0.9520	0.0242	0.0238	0.9327
	P : A	0.9047	0.0554	0.0399	0.8675

Table 10-9. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 8 Reading

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.8146	0.7429	0.5940		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.8208	0.7400		
	Nearing Proficiency	0.6237	0.5051		
	Proficient	0.7591	0.6722		
	Advanced	0.8955	0.8422		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9728	0.0130	0.0142	0.9616
	NP : P	0.9505	0.0253	0.0241	0.9306
	P : A	0.8907	0.0595	0.0498	0.8465

Table 10-10. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 10 Reading

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7786	0.6983	0.5617		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.8139	0.7309		
	Nearing Proficiency	0.6369	0.5318		
	Proficient	0.8062	0.7521		
	Advanced	0.8627	0.7254		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9435	0.0269	0.0297	0.9208
	NP : P	0.9074	0.0510	0.0416	0.8713
	P : A	0.9270	0.0521	0.0209	0.8993

Table 10-11. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 3 Mathematics

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7762	0.6969	0.5750		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.8364	0.7704		
	Nearing Proficiency	0.6016	0.4899		
	Proficient	0.7673	0.7002		
	Advanced	0.8724	0.7715		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9470	0.0265	0.0265	0.9257
	NP : P	0.9208	0.0434	0.0357	0.8899
	P : A	0.9070	0.0610	0.0319	0.8724

Table 10-12. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 4 Mathematics

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7776	0.6975	0.5675		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.8106	0.7248		
	Nearing Proficiency	0.6013	0.4891		
	Proficient	0.7770	0.7113		
	Advanced	0.8767	0.7828		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9492	0.0241	0.0268	0.9286
	NP : P	0.9187	0.0431	0.0381	0.8867
	P : A	0.9083	0.0589	0.0328	0.8731

Table 10-13. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 5 Mathematics

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7795	0.6978	0.5633		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.7909	0.6856		
	Nearing Proficiency	0.6424	0.5342		
	Proficient	0.7753	0.7103		
	Advanced	0.8804	0.7853		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9551	0.0200	0.0248	0.9367
	NP : P	0.9154	0.0444	0.0403	0.8821
	P : A	0.9084	0.0594	0.0321	0.8729

Table 10-14. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 6 Mathematics

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7734	0.6904	0.5753		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.8010	0.7186		
	Nearing Proficiency	0.6525	0.5494		
	Proficient	0.7583	0.6786		
	Advanced	0.8919	0.8049		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9376	0.0306	0.0319	0.9128
	NP : P	0.9110	0.0496	0.0394	0.8760
	P : A	0.9241	0.0491	0.0268	0.8945

Table 10-15. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 7 Mathematics

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7614	0.6766	0.5558		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.7908	0.7071		
	Nearing Proficiency	0.6104	0.5028		
	Proficient	0.7463	0.6629		
	Advanced	0.8877	0.8001		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9327	0.0334	0.0338	0.9061
	NP : P	0.9089	0.0507	0.0403	0.8731
	P : A	0.9182	0.0524	0.0294	0.8861

Table 10-16. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 8 Mathematics

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7741	0.6905	0.5769		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.7757	0.6916		
	Nearing Proficiency	0.6685	0.5673		
	Proficient	0.7597	0.6770		
	Advanced	0.8973	0.8139		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9328	0.0343	0.0329	0.9065
	NP : P	0.9134	0.0495	0.0371	0.8792
	P : A	0.9274	0.0469	0.0258	0.8988

Table 10-17. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 10 Mathematics

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7795	0.6949	0.5699		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.8258	0.7554		
	Nearing Proficiency	0.7282	0.6456		
	Proficient	0.7891	0.7090		
	Advanced	0.8425	0.6817		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9241	0.0378	0.0381	0.8937
	NP : P	0.9004	0.0592	0.0404	0.8617
	P : A	0.9549	0.0326	0.0125	0.9368

Table 10-18. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 4 Science

Accuracy and Consistency of Classification Indices					
Overall Indices	Accuracy	Consistency	Kappa (κ)		
	0.7619	0.6731	0.5062		
		Accuracy	Consistency		
Indices Conditional on Level	*Novice*	0.7360	0.5811		
	Nearing Proficiency	0.7363	0.6608		
	Proficient	0.7613	0.6981		
	Advanced	0.8400	0.6608		
		Accuracy			Consistency
Indices for Dichotomous Decisions Around Cut Points		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9645	0.0139	0.0216	0.9494
	NP : P	0.8835	0.0657	0.0508	0.8395
	P : A	0.9138	0.0654	0.0208	0.8817

Table 10-19. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 8 Science

Accuracy and Consistency of Classification Indices				
Overall Indices	Accuracy	Consistency	Kappa (κ)	
	0.7657	0.6768	0.5228	
Indices Conditional on Level		Accuracy	Consistency	
	Novice	0.7700	0.6479	
	Nearing Proficiency	0.7198	0.6377	
	Proficient	0.7790	0.7130	
	Advanced	0.8392	0.6697	
Indices for Dichotomous Decisions Around Cut Points		Accuracy		Consistency
		Accuracy	*False Positives*	*False Negatives*
	N : NP	0.9501	0.0213	0.0286	0.9296
	NP : P	0.8869	0.0636	0.0495	0.8436
	P : A	0.9286	0.0527	0.0187	0.9008

Table 10-20. 2007-08 Montana CRT: Accuracy and Consistency of Performance Level Classifications—Grade 10 Science

Accuracy and Consistency of Classification Indices				
Overall Indices	Accuracy	Consistency	Kappa (κ)	
	0.7535	0.6631	0.5405	
Indices Conditional on Level		Accuracy	Consistency	
	Novice	0.8118	0.7379	
	Nearing Proficiency	0.7272	0.6452	
	Proficient	0.6846	0.5813	
	Advanced	0.8606	0.7334	
Indices for Dichotomous Decisions Around Cut Points		Accuracy		Consistency
		Accuracy	*False Positives*	*False Negatives*
	N : NP	0.9205	0.0399	0.0396	0.8890
	NP : P	0.8992	0.0606	0.0402	0.8601
	P : A	0.9332	0.0460	0.0208	0.9070

Chapter 11. SCALING AND EQUATING

The purpose of equating is to ensure that scores obtained from different forms of a test are equivalent to each other. Equating may be used if multiple test forms are administered in the same year, as well as to equate one year's forms to those given in the previous year. Equating ensures that students are not given an unfair advantage or disadvantage because the test form given in one year is easier or harder than the form given in the other year. Once test scores for the forms are placed on an equivalent raw score scale, they then get translated, through the scaling process, to the score scale that is used for reporting. For the 2007-08 Montana CRT, equating was performed for reading and mathematics, grades 3 through 8 and 10.

A standard setting meeting was conducted for the new science tests in July 2008 (the standard setting report is included as Appendix C). Thus, operational 2007-08 data were used to set the science standards, and subsequent administrations of the Montana CRT science tests will be equated back to the 2007-08 scale. The cut scores, which were set on the θ metric and transformed into scale scores (explained in Section 11.3), will remain fixed in the future unless standards are reset for any reason.

11.1 General Rules

The following general rules are contained in the equating plan for the Montana CRT:

- The goal is to have as many items as possible on the common form constitute the equating set.
- Items used for equating cannot be altered from their appearance in the previous form in any way.
- Whenever possible, items in the equating set should be selected so that they are within three or four positions of their location on the previous form.

- Passage sets selected for equating should consist of all, or most, of the items associated with the passage.
- The equating set, as a whole group of items, should mirror the characteristics of the common form in terms of content and statistics.

To determine the final set of equating items for each grade level and subject combination, a differential item functioning (DIF) approach using the delta plot method was applied. The 2007-08 and 2006-07 p-values of each multiple-choice item were transformed to the delta metric. The delta scale is an inverse normal transformation of percentage correct to a linear scale with a mean of 13 and standard deviation of 4 (Holland & Wainer, 1993). A high delta value indicates a difficult item. For constructed-response items, the average score divided by the maximum possible score, i.e., the adjusted p-value, was transformed to the delta metric. The delta values for the potential equating items were computed for each subject in each grade level.

Once all the delta values were calculated for a particular subject and grade, a trend line was fit to the set of points. The perpendicular distance of each item to the regression line was then computed. Items that were not more than three standard deviations away from the regression line were used as equating items. As a result of the delta analyses, eight items were excluded for use as equating items, one each in the following grade-contents: Grades 3, 5, 8, and 10 reading; grades 5, 7, 8, and 10 mathematics.

11.2 IRT Equating

Equating for the Montana CRT used the *anchor-test-nonequivalent-groups* design described by Petersen, Kolen, & Hoover (1989). The FCIP (fixed common-item item parameter) method was used, in which the equating or "anchor" items from the previous year's administration were identified during this year's IRT (item response theory) calibration (explained shortly) and their parameters fixed to last year's values. This method results in all person and item parameters being

on the same θ scale as the previous year. The procedures used for equating and scaling the Montana CRT (a) do not change the rank ordering of students, (b) give more weight to particular items, or (c) change students' performance-level classifications. Note that the groups of students who took the Montana CRT in 2006-07 and 2007-08 were not equivalent. IRT is particularly useful in equating for such "nonequivalent" groups (Allen & Yen, 1979).

IRT uses mathematical models to define a relationship between an unobserved measure of student ability, usually referred to as theta (θ), and the probability (P_{jk}) of person *k* getting a dichotomous item *j* correct (or of getting a particular score on a polytomous item *j*). In IRT, it is assumed that all items are independent measures of the same construct or ability (i.e., the same θ). There are several IRT models commonly used to specify the relationship between θ and *p*. For the Montana CRT tests, the generalized partial credit model (GPCM) was used for the constructed-response items and the one-parameter logistic (1PL) model was used for multiple-choice and short-answer items.

The GPCM model can be defined as

$$P_{jk}(\theta) = \frac{\exp \sum_{v=0}^{k} \left[Da_j \left(\theta - b_j + d_v \right) \right]}{\sum_{c=1}^{m} \exp \sum_{v=1}^{c} \left[Da_j \left(\theta - b_j + d_v \right) \right]}$$

where
j indexes the items,
k indexes students,
a represents item discrimination,
b represents item difficulty,
d represents category step parameter, and
D is a normalizing constant equal to 1.701.

In the case of the 1PL model used for the Montana CRT, the a_j term in the above equation is set equal to 1.0 for all items.

For dichotomous items there are also no step parameters (d_v), so the above equation further reduces to

$$P_j(\theta) = \frac{\exp(\theta - b_j)}{1 + \exp(\theta - b_j)}$$

For more information on IRT and these models, the reader is referred to Hambleton and Swaminathan (1985).

The process of determining the specific mathematical relationship between θ and P_{jk} is referred to as item calibration. Once items are calibrated, they are defined by a set of parameters which specify a non-linear relationship between θ and P_{jk}. For more information about item calibration the reader is referred to Lord and Novick (1968) or Hambleton and Swaminathan (1985).

PARSCALE v3.5 (Muraki & Bock, 1999) software was used to do all IRT analyses for the Montana CRT tests. The item parameter files resulting from the analyses are provided in Appendix A. Each item occupied only one block in the calibration run, and the 1.701 normalizing constant was used. A default convergence criterion was set at 0.001, and all calibrations converged within 32 iterations.

11.3 Translating Raw Scores to Scaled Scores and Performance Levels

Montana CRT scores in each content area are reported on a scale that ranges from 200 to 300. Scaled scores supplement the Montana CRT performance-level results by providing information about the position of a student's results within a performance level. School- and district-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores or total number of points, on the Montana CRT tests are translated to scaled scores using a data analysis process called *scaling*. Scaling simply converts raw points from one scale to another. In the same way that distance can be expressed in miles or kilometers, or monetary value can be

expressed in terms of U.S. dollars or Canadian dollars, student scores on each Montana CRT could be expressed as raw scores (i.e., total points earned) or scaled scores. It is also important to note that the specific raw score to scale score conversion formula varies from content area to content area within grade, and between grades as well. For example, the scaling conversion formula for Montana's Grade 4 reading test differs from that of the Grade 4 mathematics test or the Grade 8 reading test.

It is important to note that converting from raw scores to scaled scores does not change the students' performance-level classifications. Given the relative simplicity of raw scores, it is fair to ask why scaled scores are used in Montana CRT reports instead of raw scores. Foremost, scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Because the standard-setting process typically results in different cut scores across content areas on a raw score basis, it is useful to transform these raw cut scores to a scale that is more easily interpretable and consistent. For the Montana CRT, a score of 225 is the cut score between the *Novice* and *Nearing Proficiency* performance levels. This is true regardless of which content area, grade, or year one may be concerned with. If one were to use raw scores, the raw cut score between *Novice* and *Nearing Proficiency* may be, for example, 35 in mathematics at grade 8, but may be 33 in mathematics at grade 10. Using scaled scores standardizes the scale one uses to interpret student performance.

Cut points for reading and mathematics tests for the Montana CRT were set at standard setting meetings held in June and July, 2006 and those for science in June, 2008. Cut points were established on the raw score scale, and these raw score cuts were used to determine the scaling coefficients for calculating the scores used for reporting (see description below and Appendix C). Cut points were also determined on the θ-scale. For scaling in 2007-08, raw score equivalents for these θ-scale cut points were determined using the test characteristic curve (TCC), and these 2007-08 raw cuts were used to calculate transformation constants.

As previously stated, student scores on the Montana CRT are reported in integer values from 200 to 300 with three scores representing cut scores on each assessment. Two of the three cut points (*Novice*/*Nearing Proficiency* and *Nearing Proficiency*/*Proficient*) are pre-set at 225 and 250, respectively; the third cut point, between *Proficient* and *Advanced*, is allowed to vary across tests, depending on where the raw score cuts were placed. Allowing the upper cut to float results in a single conversion equation for each test. Table 11-1 presents the scaled score range for each performance level in each grade-content area combination.

Table 11-1. 2007-08 Montana CRT: Scaled Score Range for each Performance Level

Grade	*Content Area*	Novice	Nearing Proficiency	Proficient	Advanced
3	Reading	200–224	225–249	250–286	287–300
	Mathematics	200–224	225–249	250–289	290–300
4	Reading	200–224	225–249	250–288	289–300
	Mathematics	200–224	225–249	250–290	291–300
	Science	200–224	225–249	250–280	281-300
5	Reading	200–224	225–249	250–286	287–300
	Mathematics	200–224	225–249	250–288	289–300
6	Reading	200–224	225–249	250–288	289–300
	Mathematics	200–224	225–249	250–286	287–300
7	Reading	200–224	225–249	250–287	288–300
	Mathematics	200–224	225–249	250–288	289–300
8	Reading	200–224	225–249	250–288	289–300
	Mathematics	200–224	225–249	250–282	283–300
	Science	200–224	225–249	250–282	283–300
10	Reading	200–224	225–249	250–288	289–300
	Mathematics	200–224	225–249	250–280	281–300
	Science	200–224	225–249	250–268	269–300

The scaled scores are obtained by a simple linear transformation of the raw scores using the fixed scaled score values noted above (225 and 250) and the associated 2007-08 raw score cut points.

The scaling coefficients were calculated using the following formula for the slope (*m*) of scaled scores as a function of raw scores.

$$m = \frac{225 - 250}{x_1 - x_2}$$

Where:
x_1 is the raw cut score for the *Novice*/*Nearing Proficiency* cut,
x_2 is the raw cut score for the *Nearing Proficiency*/*Proficient* cut

In other words, the slope is the ratio between the scale score and raw score differences at the fixed cut points.

The intercept (*b)* of the function is found either by

$$b = 225 - m(x_1) \text{ or}$$
$$b = 250 - m(x_2)$$

and represents the resultant scale score if, at the rate of the slope, the raw score fell from one of the cut points to zero.

Scaled scores were then calculated using the resulting linear function:

$$ss = m(x) + b$$

where
x represents a student's raw score.

The values obtained using this formula were rounded to the nearest integer and truncated, as necessary, such that no student received a score below 200 or higher than 300. Additional information regarding raw scores, scaled scores, performance level descriptors, and content-specific descriptors may be found in Appendix D.

Chapter 12. REPORTING

The Montana CRT tests were designed to measure student performance against Montana's content standards. Consistent with this purpose, results on the CRT were reported in terms of performance levels that describe student performance in relation to these established state standards. There are four performance levels: *Advanced*, *Proficient*, *Nearing Proficiency*, and *Novice*. (The CRT Performance Level Descriptors are given in Appendix D as are student distributions within the raw and scaled score ranges of the performance levels.) Students receive a separate performance-level classification (based on total scaled score) in each content area.

State results were provided to OPI via a secure Web site. Reading and mathematics reporting data for the 2007-08 Montana CRT were made available to systems and schools online via the Montana Analysis and Reporting System (MARS) on June 10, 2008; science results for grade 4, 8 and 10 on September 2, 2008. Student Reports were delivered to schools on September 18, 2008. System Test Coordinators and teachers were also provided with copies of the *Guide to Interpreting the 2007 Criterion-Referenced Test and CRT-Alternate Assessment Reports* to assist them in understanding the connection between the assessment and the classroom. The guide provides information about the assessment and the use of assessment results.

School- and system-level results are reported as the number and percentage of students attaining each performance level at each grade level tested. "Decision Rules" were formulated in early 2008 by OPI and Measured Progress to identify students who, during the reporting process, were to be excluded from school and system-level reports. A copy of these "Decision Rules" is included in this report as Appendix F. Disaggregations of students are also reported at the school and system levels. The CRT reports include:

- Student Reports (paper);
- Class Roster & Item-Level Reports (online/interactive);
- School Summary Reports (online/pdf); and
- System Summary Reports (online/pdf).

Sample reports are included as Appendix E.

12.1 Montana Analysis and Reporting System (MARS)

After a year of gather input and feedback on the analysis and reporting system from Montana system administrators and principals the introduction of MARS in Montana has been a huge success. Measured Progress's system administrator reports that the site is accessed on an average of 10 times a day since its release in early June and due to the intuitive design of the system helps desk calls and training requests have been minimal.

Using advanced Web technology, *MARS* gives Montana educators and administrators the ability to filter data based on test year, grade level, content area, standard, and student subgroup. This allows administrators to isolate cross-sections of the results and identify areas of strong or poor performance.

The confidential nature of the data in *MARS* necessitates the strict enforcement of site security. All transmissions are done over Secure Socket Layers (SSL). A system of user role definitions and permissions dictates the scope of access granted to individual users. Organizations (system or school levels) are given administrative power to grant or deny access to their data within the system, and have the ability to disable users. Personnel using *MARS* may be granted permission to view students' results at an organizational level, or only a select group as defined by the administrator. Predefined reports are included in the system, as is the ability to render and print additional copies.

Chapter 13. VALIDITY SUMMARY

As stated in the overview chapter, the *Standards for Educational and Psychological Testing* (AERA, et al., 1999) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. The evidence sources around test content, response processes, internal structure, relationship to other variables, and consequences of testing speak to different *aspects* of validity but are not distinct *types* of validity. Instead, each of these contributes to a body of evidence about the comprehensive validity of score interpretations.

Evidence on test content validity is meant to determine how well the assessment tasks represent the curriculum and standards for each subject and grade level. Content validation is informed by the item development process, including how the test blueprints and test items align to the curriculum and standards. Viewed through this lens provided by the *Standards*, evidence based on test content was extensively described in Chapters 2 through 6. Item alignment with Montana content standards; item bias, sensitivity and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training are all components of validity evidence based on test content. As discussed, all CRT test questions are aligned by Montana educators to specific Montana Content Standards and undergo several rounds of review for content fidelity and appropriateness. Items are presented to students in multiple formats (constructed-response, short-answer and multiple-choice). Finally, tests are administered according to state-mandated standardized procedures, with allowable accommodations, and all test proctors are required to attend annual training sessions.

The scoring information in Chapter 8 described the steps taken to train and monitor hand-scorers, as well as quality control procedures related to scanning and machine scoring. To speak to

student response processes, however, additional studies would be helpful and might include an investigation of students' cognitive methods using think-aloud protocols.

Evidence based on internal structure was presented in great detail in the discussions of item properties, scale dimensionality, test reliability, and scaling and equating in Chapters 9 through 11. Technical characteristics of the internal structure of the assessments were presented in terms of classical item statistics (item difficulty, item-test correlation), differential item functioning analyses, tests of dimensionality and computation of dimensionality effect sizes, a variety of reliability coefficients, standard errors of measurement, and item response theory parameters and procedures. It was explained how each test is equated to the same grade and content test from the prior year in order to preserve the meaning of scores over time. It was shown that item difficulty and discrimination indices were in acceptable and expected ranges. The degree of multidimensionality detected in each grade-content was reported to be too small to warrant further inquiry. And finally, all tests exhibited not only industry standard levels of reliability for large-scale assessments, but accurate and consistent classification decisions as well.

Evidence based on the consequences of testing was addressed in the scaled scores and reporting information found in Chapters 11 and 12, respectively (as well as in the test interpretation guide, which is a separate document that is referenced in the discussion of reporting). the information contained therein spoke to the efforts undertaken to promote accurate and clear information provided to the public regarding test scores. The advantages of using scaled scores and performance levels for reporting results across content areas, grade levels, and subsequent years was discussed. The several different standard reports provided to stakeholders were described and examples were shown. It may be mentioned here as well that a data analysis tool is provided to each school system that allows educators the flexibility to customize reports for local needs. Additional evidence of the consequences of testing could be supplemented with broader investigation of the impact of testing on student learning.

To further support the validation of the assessment program, additional studies might be considered to provide evidence regarding the relationship of CRT results to other variables include the extent to which scores from the CRT assessments converge with other measures of similar constructs, and the extent to which they diverge from measures of different constructs. Relationships among measures of the same or similar constructs can sharpen the meaning of scores and appropriate interpretations by refining the definition of the construct.

The evidence presented in this report supports inferences of student achievement on the content represented on the Montana Content Standards for reading, mathematics, and science for the purposes of program and instructional improvement and as a component of school accountability.

SECTION IV—REFERENCES

Allen, Mary J. & Yen, Wendy M. (1979). *Introduction to Measurement Theory*. Belmont, CA: Wadsworth, Inc.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bock, R. D., and E. Muraki (1999). *PARSCALE: Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software.

Brown, F. G. (1983). *Principles of Educational and Psychological Testing* 3rd ed. Fort Worth, TX: Holt, Rinehart, and Winston.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Dorans, N. J., and P. W. Holland (1993). DIF detection and description. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* pp. 35–66. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., and E. Kulick (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). New York: John Wiley & Sons, Inc

Hambleton, R. K., and W. J. van der Linden (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

Hambleton, R. K., and H. Swaminathan (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer Academic Publishers.

Joint Committee on Testing Practices (2004). *Code of Fair Testing Practices in Education*. Washington, DC: American Psychological Association. Available for download at **http://www.apa.org/science/fairtestcode.html**.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

Lord, F.M., and M. R. Novick (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). *Scaling, Norming, and Equating*. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 221-262).

APPENDICES

APPENDIX A—ITEM PARAMETER FILES

Table A-1. 2007-08 MT CRT: Item Parameter Files: Grade 3 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
43120	1	0	-1.75842	0				
42971	1	0	-0.6653	0				
43145	1	0	-1.25634	0				
43104	1	0	-0.11351	0				
42967	1	0	-0.53053	0				
43022	1	0	-0.57591	0				
43131	1	0	-1.40196	0				
42978	1	0	0.26468	0				
43106	1	0	-1.42915	0				
43094	1	0	-0.23772	0				
42982	1	0	-0.35522	0				
43012	1	0	-1.01823	0				
43024	1	0	-0.16055	0				
43018	1	0	-0.35377	0				
42962	1	0	-0.671	0				
42984	1	0	-0.05846	0				
42980	1	0	-0.50318	0				
43165	1	0	-0.88618	0				
43114	1	0	-1.26405	0				
43148	1	0	-1.03741	0				
43082	1	0	-0.59862	0				
42983	1	0	-0.63564	0				
43130	1	0	-0.13994	0				
42960	1	0	-0.68989	0				
42965	1	0	-0.79166	0				
43103	1	0	-0.2594	0				
43078	1	0	-0.44516	0				
42956	1	0	-0.97451	0				
43013	1	0	0.26734	0				
43002	1	0	-1.21887	0				
42986	1	0	-0.57217	0				
43110	1	0	-0.50318	0				
43020	1	0	-0.38263	0				
42977	1	0	-1.11393	0				
42979	1	0	-0.18049	0				
43154	1	0	0.16831	0				
34689	1	0	-1.71202	0				
43014	1	0	-1.89661	0				
43004	1	0	-0.87468	0				
43064	1	0	-1.08226	0				
43162	1	0	-0.61731	0				
42989	1	0	0.06025	0				
242748	1	0	-1.21528	0				
43087	1	0	-0.8511	0				
42975	1	0	0.2225	0				
43102	1	0	-1.46424	0				
43108	1	0	-0.35929	0				
42990	1	0	-0.34537	0				
43136	1	0	-0.69515	0				
43105	1	0	-0.40321	0				
43009	1	0	-0.35697	0				

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
43015	1	0	-0.00455	0				
43071	1	0	-0.73059	0				
43008	1	0	-0.33699	0				
42952	1	0	-1.0773	0				
43028	1	0	-0.48605	0				
42994	1	0	-0.41831	0				
42992	1	0	-0.88309	0				
43096	1	0	-0.50678	0	0.44976	0.19812	0.14843	-0.79631
42996	1	0	0.05487	0	-0.26306	0.8443	-0.4737	-0.10754

Table A-1. 2007-08 MT CRT: Item Parameter Files: Grade 4 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
43377	1	0	-0.61204	0				
43349	1	0	-1.04623	0				
43142	1	0	0.31477	0				
43340	1	0	-0.18907	0				
43300	1	0	0.50792	0				
43288	1	0	-0.40564	0				
43306	1	0	-0.07841	0				
43338	1	0	0.41154	0				
244335	1	0	-0.03923	0				
43156	1	0	0.4094	0				
34581	1	0	-0.39681	0				
43344	1	0	0.40914	0				
43184	1	0	-0.15538	0				
43324	1	0	-0.52358	0				
43194	1	0	0.1071	0				
43361	1	0	-0.44476	0				
242880	1	0	0.07196	0				
43357	1	0	-1.01434	0				
43355	1	0	-1.13862	0				
43322	1	0	-0.7526	0				
43197	1	0	-0.57264	0				
43367	1	0	-0.17452	0				
243119	1	0	0.29783	0				
43342	1	0	-0.71616	0				
43164	1	0	-0.24192	0				
43310	1	0	-0.56228	0				
43201	1	0	-0.08621	0				
43272	1	0	0.36941	0				
43302	1	0	-0.59528	0				
34778	1	0	0.69699	0				
43291	1	0	-0.26176	0				
43169	1	0	-0.22512	0				
43191	1	0	0.14646	0				
43328	1	0	-0.56172	0				
43182	1	0	0.15444	0				
243063	1	0	1.12863	0				
43353	1	0	-0.95109	0				
43363	1	0	-0.99137	0				
43253	1	0	-0.83241	0				
244304	1	0	-0.23539	0				
43298	1	0	0.15444	0				
247987	1	0	0.903	0				
43334	1	0	-0.40804	0				
43386	1	0	0.32302	0				
43241	1	0	-0.45909	0				
43282	1	0	0.17055	0				
35198	1	0	0.63013	0				
43326	1	0	-0.30206	0				
43244	1	0	0.2392	0				
35217	1	0	0.3678	0				
43336	1	0	-0.25147	0				
35206	1	0	0.24672	0				
43160	1	0	-0.53908	0				
43167	1	0	-0.01707	0				
43276	1	0	-0.56695	0				
43193	1	0	0.60127	0				
43187	1	0	-0.31134	0				
243180	1	0	0.50238	0				
246631	1	0	0.16057	0	0.61796	0.14665	0.12198	-0.88659
35788	1	0	0.38018	0	0.59612	0.62539	-0.55546	-0.66605

Table A-3. 2007-08 MT CRT: Item Parameter Files: Grade 5 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
43568	1	0	-1.61836	0				
43477	1	0	-0.36273	0				
43408	1	0	-0.67344	0				
43566	1	0	-0.38491	0				
43478	1	0	-0.82003	0				
43574	1	0	-0.55302	0				
43514	1	0	0.28713	0				
43451	1	0	-0.39024	0				
43437	1	0	-0.40699	0				
43558	1	0	-0.76811	0				
43469	1	0	-0.03110	0				
237100	1	0	-0.58900	0				
43429	1	0	0.10156	0				
43532	1	0	0.52259	0				
43535	1	0	-0.26243	0				
43526	1	0	0.56000	0				
43530	1	0	-0.00023	0				
43520	1	0	-0.54743	0				
43473	1	0	-1.10262	0				
43564	1	0	-0.75655	0				
43419	1	0	-0.73819	0				
43525	1	0	-0.38251	0				
43581	1	0	-0.22527	0				
43517	1	0	-0.47306	0				
43559	1	0	-0.70546	0				
43453	1	0	0.00177	0				
43435	1	0	-0.67980	0				
43443	1	0	-0.26819	0				
43528	1	0	0.15757	0				
243040	1	0	0.45952	0				
43510	1	0	1.12373	0				
43518	1	0	0.64470	0				
43521	1	0	-0.82331	0				
43417	1	0	-0.43752	0				
43504	1	0	-0.60379	0				
43534	1	0	0.09159	0				
34517	1	0	-0.83964	0				
43445	1	0	-1.92720	0				
43498	1	0	-0.91554	0				
43500	1	0	-0.32755	0				
236242	1	0	-0.18369	0				
43433	1	0	-0.08805	0				
43480	1	0	-0.23802	0				
43457	1	0	0.53983	0				
43556	1	0	-0.46846	0				
43502	1	0	-0.10403	0				
43431	1	0	0.17210	0				
34658	1	0	0.25118	0				
43543	1	0	-0.05492	0				
43486	1	0	-0.15615	0				
43411	1	0	-0.05497	0				
43484	1	0	0.47881	0				
43524	1	0	-0.55566	0				
43413	1	0	-0.06794	0				
43421	1	0	-0.72319	0				
43552	1	0	-0.00810	0				
43563	1	0	-1.18071	0				
43544	1	0	-0.05974	0				
43594	1	0	0.28866	0	0.47468	0.48378	0.01024	-0.96871
242957	1	0	-0.33415	0	0.16109	0.46457	-0.61327	-0.01239

Table A-4. 2007-08 MT CRT: Item Parameter Files: Grade 6 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
43870	1	0	-1.22295	0				
43861	1	0	-0.72934	0				
43910	1	0	-0.54485	0				
43874	1	0	0.07925	0				
43854	1	0	-0.87483	0				
43930	1	0	0.16565	0				
43924	1	0	-0.09756	0				
43897	1	0	0.16754	0				
43912	1	0	0.05001	0				
43975	1	0	-0.18225	0				
43927	1	0	0.52649	0				
43993	1	0	0.11614	0				
43946	1	0	-0.4768	0				
43852	1	0	-1.08033	0				
44039	1	0	-1.5969	0				
44070	1	0	-0.64646	0				
44064	1	0	-0.49936	0				
44040	1	0	-0.18507	0				
43956	1	0	-0.21688	0				
43447	1	0	-0.54967	0				
43879	1	0	0.20761	0				
44074	1	0	0.11473	0				
44062	1	0	0.33578	0				
44037	1	0	0.66953	0				
43949	1	0	0.18127	0				
44015	1	0	-0.26434	0				
43963	1	0	-0.52545	0				
43868	1	0	-0.14489	0				
34913	1	0	0.11366	0				
44001	1	0	0.00562	0				
43997	1	0	-0.64611	0				
44033	1	0	0.35998	0				
44060	1	0	-0.50804	0				
43977	1	0	-0.10183	0				
44094	1	0	-1.05269	0				
44066	1	0	-0.90571	0				
43893	1	0	-0.79609	0				
43953	1	0	-0.51433	0				
44004	1	0	-0.70185	0				
44080	1	0	-0.73818	0				
43459	1	0	-0.18232	0				
44059	1	0	0.53229	0				
43939	1	0	0.20342	0				
43887	1	0	-0.19765	0				
43847	1	0	-0.60595	0				
44072	1	0	0.54315	0				
43966	1	0	-0.21851	0				
44027	1	0	0.37807	0				
44021	1	0	0.11713	0				
43981	1	0	-0.31569	0				
44044	1	0	0.88191	0				
43995	1	0	-0.83918	0				
34539	1	0	0.67864	0				
43488	1	0	-0.1801	0				
43973	1	0	-0.86807	0				
43968	1	0	-0.15576	0				
43907	1	0	0.71447	0				
43916	1	0	0.05329	0				
43989	1	0	0.44022	0	0.10934	0.04367	-0.43107	0.27806
236926	1	0	-0.18074	0	0.5525	-0.21847	-0.21633	-0.11771

Table A-5. 2007-08 MT CRT: Item Parameter Files: Grade 7 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
43672	1	0	-0.09391	0				
43832	1	0	-1.05834	0				
43883	1	0	-0.20927	0				
43836	1	0	-0.26535	0				
43820	1	0	0.75851	0				
43875	1	0	-0.40392	0				
43846	1	0	-0.03193	0				
43772	1	0	0.56658	0				
43796	1	0	0.2368	0				
43689	1	0	-0.7307	0				
43839	1	0	-0.04973	0				
43865	1	0	0.31574	0				
43871	1	0	-0.09969	0				
44190	1	0	-0.02011	0				
43780	1	0	-0.73565	0				
43651	1	0	-0.22133	0				
43714	1	0	-0.42757	0				
43809	1	0	-0.03305	0				
43787	1	0	-0.2956	0				
43685	1	0	0.43535	0				
43657	1	0	0.07293	0				
43782	1	0	0.14336	0				
43721	1	0	-0.01253	0				
43663	1	0	0.04344	0				
43701	1	0	0.13993	0				
43746	1	0	0.52524	0				
43693	1	0	-0.20666	0				
43715	1	0	0.72378	0				
43698	1	0	0.68237	0				
43671	1	0	0.41036	0				
43659	1	0	0.05437	0				
43771	1	0	-0.61974	0				
43753	1	0	-0.17651	0				
43666	1	0	-0.63741	0				
43777	1	0	-0.56046	0				
43695	1	0	-0.78138	0				
43805	1	0	0.92078	0				
43700	1	0	-0.03531	0				
43645	1	0	0.37471	0				
43735	1	0	-0.96908	0				
43817	1	0	-0.20742	0				
43750	1	0	-0.0552	0				
44238	1	0	0.32489	0				
43731	1	0	0.50346	0				
43885	1	0	-0.04532	0				
43646	1	0	0.02663	0				
43668	1	0	0.48687	0				
43675	1	0	-0.01107	0				
43711	1	0	-0.10855	0				
43763	1	0	-0.29733	0				
43654	1	0	-0.45415	0				
43788	1	0	-0.49993	0				
44156	1	0	0.50684	0				
44211	1	0	-0.12585	0				
43719	1	0	-0.08218	0				
43799	1	0	-0.40464	0				
43909	1	0	0.53572	0				
43900	1	0	0.2184	0				
43914	1	0	0.56056	0	0.5457	0.46558	-0.1529	-0.85839
43829	1	0	0.36452	0	0.03857	-0.35776	1.05146	-0.73227

Table A-6. 2007-08 MT CRT: Item Parameter Files: Grade 8 Math

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
44201	1	0	-0.60196	0				
44183	1	0	-0.63717	0				
44207	1	0	-0.06197	0				
44621	1	0	-0.34167	0				
44209	1	0	-0.38185	0				
44176	1	0	0.58136	0				
44253	1	0	0.18846	0				
43840	1	0	0.53935	0				
44179	1	0	0.34594	0				
44255	1	0	0.02911	0				
244557	1	0	-0.52706	0				
43824	1	0	0.15601	0				
44177	1	0	0.38488	0				
44189	1	0	0.19806	0				
44151	1	0	0.21076	0				
244502	1	0	0.25367	0				
44210	1	0	-0.26086	0				
43888	1	0	0.0351	0				
44188	1	0	-0.22552	0				
43744	1	0	0.33586	0				
44143	1	0	0.06842	0				
44116	1	0	-0.54472	0				
44153	1	0	-0.57473	0				
44224	1	0	0.11549	0				
44648	1	0	-0.12312	0				
44161	1	0	0.61608	0				
44184	1	0	0.40833	0				
44245	1	0	0.29899	0				
44256	1	0	0.69727	0				
44145	1	0	-0.40372	0				
44244	1	0	-0.23381	0				
44236	1	0	-0.05191	0				
44130	1	0	-0.24874	0				
44220	1	0	-0.11116	0				
44234	1	0	0.15022	0				
44168	1	0	-0.62284	0				
44232	1	0	0.39488	0				
44186	1	0	-0.74477	0				
44123	1	0	0.19388	0				
44239	1	0	-0.85075	0				
244552	1	0	-0.44467	0				
44127	1	0	-0.81125	0				
44205	1	0	-0.73388	0				
44227	1	0	0.20921	0				
44243	1	0	-0.2123	0				
243343	1	0	-0.22849	0				
44140	1	0	0.24932	0				
244493	1	0	-1.11662	0				
244622	1	0	-0.28792	0				
44099	1	0	-0.66959	0				
243315	1	0	0.29539	0				
244528	1	0	0.17657	0				
44141	1	0	-0.81254	0				
44154	1	0	0.04347	0				
44149	1	0	-0.47353	0				
44199	1	0	0.04378	0				
44191	1	0	0.52013	0				
44121	1	0	-0.04209	0				
34999	4	0	0.30142	0	-0.0610	0.4883	-0.1283	-0.2990
248854	4	0	0.26691	0	-0.12113	1.02194	-0.1918	-0.70902

Table A-7. 2007-08 MT CRT: Item Parameter Files: Grade 10 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
43606	1	0	-0.12688	0				
43725	1	0	0.74114	0				
43608	1	0	0.02003	0				
43797	1	0	0.47572	0				
43614	1	0	-0.56073	0				
43800	1	0	-0.02531	0				
43884	1	0	0.77729	0				
43943	1	0	-0.31022	0				
43712	1	0	0.44784	0				
43616	1	0	0.64194	0				
166916	1	0	-0.06721	0				
43665	1	0	-0.16744	0				
43611	1	0	0.21619	0				
43703	1	0	0.5389	0				
43889	1	0	-0.30105	0				
44024	1	0	-0.17602	0				
43951	1	0	0.586	0				
43717	1	0	-0.60682	0				
43661	1	0	-0.32508	0				
43841	1	0	0.82185	0				
43926	1	0	0.15036	0				
43969	1	0	-1.24676	0				
43819	1	0	-0.1176	0				
43844	1	0	0.48987	0				
43964	1	0	-0.25428	0				
43728	1	0	0.9968	0				
35232	1	0	-0.00326	0				
43638	1	0	0.24502	0				
43830	1	0	0.31378	0				
43948	1	0	0.17974	0				
43789	1	0	-0.44705	0				
43617	1	0	0.09183	0				
43697	1	0	-0.24901	0				
43833	1	0	0.20131	0				
44573	1	0	0.05686	0				
43880	1	0	-0.56929	0				
43765	1	0	-0.01581	0				
43740	1	0	0.20961	0				
43729	1	0	-0.1639	0				
43807	1	0	-0.18235	0				
43609	1	0	0.18788	0				
43628	1	0	0.121	0				
43803	1	0	-0.13899	0				
43785	1	0	0.14184	0				
43959	1	0	0.54213	0				
43629	1	0	-0.26633	0				
242987	1	0	-0.68182	0				
43877	1	0	0.39238	0				
43636	1	0	-0.10672	0				
43837	1	0	-0.86	0				
43633	1	0	0.687	0				
34639	1	0	0.27949	0				
34853	1	0	0.47378	0				
43826	1	0	0.2654	0				
43710	1	0	-0.04172	0				
43674	1	0	0.09657	0				
43895	1	0	-0.19897	0				
43670	1	0	0.42016	0				
43643	1	0	0.59838	0	0.4238	0.1754	-0.0175	-0.5817
43872	1	0	0.3308	0	-0.4741	0.8748	-0.6141	0.2135

Table A-8. 2007-08 MT CRT: Item Parameter Files: Grade 3 Reading

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
244235	1	0	-0.65492	0				
247940	1	0	-0.86195	0				
247847	1	0	-1.30167	0.01917				
247849	1	0	-1.05183	0				
247850	1	0	-0.88532	0				
247954	1	0	-0.84357	0				
243252	1	0	-0.82542	0				
42896	1	0	-0.56931	0				
42897	1	0	0.21287	0				
42899	1	0	-0.06854	0				
42900	1	0	-0.28022	0				
42907	1	0	0.37982	0				
42904	1	0	-0.27767	0				
42903	1	0	-0.48678	0				
42906	1	0	-0.37602	0				
44735	1	0	-0.49373	0				
42910	1	0	0.15262	0				
42908	1	0	-1.04518	0				
42912	1	0	-1.72304	0				
42441	1	0	-1.01943	0				
42444	1	0	-1.29315	0				
42446	1	0	-0.89661	0				
42455	1	0	-0.52213	0				
42457	1	0	-1.81744	0				
44644	1	0	-0.67431	0				
42463	1	0	0.63535	0				
42833	1	0	-0.03925	0				
42834	1	0	0.02036	0				
42839	1	0	-0.25476	0				
42835	1	0	0.80713	0				
42837	1	0	-0.77253	0				
42838	1	0	-0.95851	0				
42840	1	0	-0.428	0				
42727	1	0	-0.54906	0				
42729	1	0	-0.03146	0				
42732	1	0	-0.39483	0				
42735	1	0	-0.33876	0				
42739	1	0	0.00141	0				
42738	1	0	-0.17595	0				
42745	1	0	-0.31917	0				
42573	1	0	-1.62439	0				
42576	1	0	-0.66116	0				
42589	1	0	-1.05285	0				
42596	1	0	-1.31597	0				
42608	1	0	-0.52608	0				
42593	1	0	-0.82543	0				
42606	1	0	-0.88546	0				
42611	1	0	-0.12824	0				
42603	1	0	-1.0295	0				
42616	1	0	-0.71464	0				
42639	1	0	-0.27015	0				
42648	1	0	-0.30536	0				
42913	4	1	0.20256	0	1.7209	0.2151	-0.7222	-1.2139
42653	4	1	0.23842	0	1.5432	0.3376	-0.4741	-1.4067

Table A-9. 2007-08 MT CRT: Item Parameter Files: Grade 4 Reading

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
244384	1	0	0.0124	0				
235578	1	0	-0.9839	0				
235579	1	0	-0.3201	0				
235583	1	0	-0.6321	0				
235585	1	0	-0.0464	0				
235587	1	0	0.0560	0				
235591	1	0	0.2785	0				
40973	1	0	0.2330	0				
40974	1	0	0.2005	0				
40975	1	0	-0.0130	0				
40982	1	0	-0.3182	0				
40977	1	0	-0.5132	0				
40979	1	0	0.3632	0				
40980	1	0	0.5066	0				
40986	1	0	0.3505	0				
40987	1	0	0.3823	0				
40983	1	0	-0.1933	0				
40985	1	0	-0.3942	0				
40990	1	0	-0.1340	0				
244303	1	0	0.4843	0				
235774	1	0	-0.2075	0				
244348	1	0	-0.5395	0				
235777	1	0	-0.8490	0				
244354	1	0	0.2320	0				
244353	1	0	-0.5214	0				
248070	1	0	-0.1535	0				
41026	1	0	0.6207	0				
41029	1	0	-0.6760	0				
41032	1	0	-0.4102	0				
41033	1	0	-0.0602	0				
41030	1	0	-0.4720	0				
41037	1	0	0.0932	0				
41038	1	0	-0.3435	0				
41137	1	0	0.3121	0				
41138	1	0	-0.6983	0				
41143	1	0	-0.6144	0				
41142	1	0	-0.2354	0				
41145	1	0	-0.8880	0				
41141	1	0	0.3656	0				
41148	1	0	-0.7969	0				
235600	1	0	-0.5035	0				
235606	1	0	-0.2397	0				
235618	1	0	-0.6671	0				
235621	1	0	-0.2644	0				
235627	1	0	0.3207	0				
244322	1	0	0.0382	0				
244357	1	0	-0.1330	0				
235640	1	0	-0.7582	0				
244324	1	0	0.2820	0				
235646	1	0	-0.3143	0				
235648	1	0	-0.3629	0				
246673	1	0	-0.0803	0				
40992	4	1	0.5428	0	1.6992	0.3193	-0.8259	-1.1926
235654	4	1	0.5016	0	1.3791	0.4225	-0.7504	-1.0512

Table A-10. 2007-08 MT CRT: Item Parameter Files: Grade 5 Reading

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
41378	1	0	-0.60506	0				
41381	1	0	-0.84836	0				
41384	1	0	-1.12432	0				
41385	1	0	-1.35256	0				
50161	1	0	-0.88691	0				
41388	1	0	-1.17235	0				
41390	1	0	0.47432	0				
41486	1	0	-0.66433	0				
41485	1	0	-0.0692	0				
41491	1	0	-0.65821	0				
41496	1	0	-0.53661	0				
41488	1	0	-1.103	0				
41512	1	0	-0.1654	0				
41499	1	0	0.21473	0				
41493	1	0	-0.88666	0				
41507	1	0	-0.22414	0				
41505	1	0	-0.35358	0				
41513	1	0	-0.3591	0				
41514	1	0	-0.75051	0				
41551	1	0	0.24093	0				
41555	1	0	0.52101	0				
41557	1	0	-0.97126	0				
41556	1	0	-0.02592	0				
41559	1	0	-1.05279	0				
41558	1	0	0.37584	0				
41562	1	0	-0.16985	0				
41396	1	0	-0.64969	0				
41398	1	0	-0.67882	0				
41402	1	0	-0.8706	0				
41400	1	0	0.18089	0				
41403	1	0	-1.23273	0				
41404	1	0	-1.17806	0				
41405	1	0	-0.95981	0				
41471	1	0	-0.69089	0				
41472	1	0	0.16944	0				
41473	1	0	-0.47243	0				
41474	1	0	-1.02631	0				
41475	1	0	-0.84578	0				
41478	1	0	-0.8851	0				
41476	1	0	-1.14192	0				
41441	1	0	-0.27929	0				
41444	1	0	-0.44356	0				
41447	1	0	-0.31953	0				
41451	1	0	-0.93374	0				
41454	1	0	-0.82346	0				
41455	1	0	-0.41289	0				
41457	1	0	-0.41239	0				
41458	1	0	0.05706	0				
41459	1	0	-0.70207	0				
41462	1	0	-0.8461	0				
41464	1	0	-1.11516	0				
41465	1	0	-0.43829	0				
41517	4	1	0.7882	0	1.2045	0.3565	-0.4026	-1.1584
41467	4	1	0.5066	0	1.6777	0.3299	-0.8323	-1.1754

Table A-11. 2007-08 MT CRT: Item Parameter Files: Grade 6 Reading

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
42350	1	0	-0.7084	0				
42352	1	0	-1.30064	0				
42360	1	0	-0.43564	0				
42357	1	0	-0.6352	0				
42358	1	0	-1.03761	0				
42355	1	0	-0.56474	0				
74008	1	0	-0.98729	0				
254034	1	0	-0.00181	0				
246594	1	0	-0.6861	0				
238628	1	0	-0.22736	0				
238669	1	0	-0.71413	0				
246595	1	0	-1.11765	0				
254020	1	0	-0.20492	0				
238650	1	0	-0.84536	0				
254101	1	0	-0.55594	0				
254019	1	0	0.06504	0				
254021	1	0	-0.48444	0				
238614	1	0	-0.50958	0				
74028	1	0	-0.30373	0				
44716	1	0	-0.37592	0				
44717	1	0	-0.7173	0				
44718	1	0	-0.12127	0				
44720	1	0	-0.4232	0				
44721	1	0	-0.17677	0				
44724	1	0	-0.26961	0				
44726	1	0	-0.85444	0				
41546	1	0	-0.0345	0				
41553	1	0	0.27426	0				
41549	1	0	-1.17873	0				
41561	1	0	0.18549	0				
41563	1	0	-0.39877	0				
41564	1	0	-0.59308	0				
41566	1	0	-0.34559	0				
41747	1	0	0.3536	0				
41749	1	0	0.24729	0				
41753	1	0	-0.97331	0				
41755	1	0	-0.943	0				
41756	1	0	-0.31237	0				
41758	1	0	-0.43496	0				
41759	1	0	-0.62859	0				
42022	1	0	-0.49739	0				
42031	1	0	-0.68437	0				
42033	1	0	-0.0844	0				
42035	1	0	-0.80234	0				
42038	1	0	-0.53021	0				
42037	1	0	-0.59772	0				
42041	1	0	-0.4066	0				
42039	1	0	-0.36976	0				
42030	1	0	-0.27558	0				
42045	1	0	-1.21661	0				
42043	1	0	-0.54898	0				
42046	1	0	0.08929	0				
239248	4	1	0.4184	0	1.6510	0.4871	-0.7365	-1.4017
42055	4	1	0.3484	0	1.4565	0.3498	-0.6810	-1.1253

Table A-12. 2007-08 MT CRT: Item Parameter Files: Grade 7 Reading

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
41769	1	0	-0.3554	0				
41771	1	0	-0.5651	0				
41786	1	0	-0.9482	0				
41788	1	0	-0.4954	0				
41791	1	0	-0.8478	0				
41794	1	0	-0.3884	0				
41795	1	0	-0.3695	0				
41892	1	0	-0.5799	0				
41894	1	0	-0.4682	0				
41895	1	0	-0.2122	0				
41896	1	0	0.1748	0				
41898	1	0	-0.5563	0				
41899	1	0	-0.4045	0				
41902	1	0	-0.3608	0				
41904	1	0	-0.3974	0				
41905	1	0	-0.0878	0				
41906	1	0	-0.0500	0				
41909	1	0	-0.7139	0				
41911	1	0	-0.8641	0				
41735	1	0	-0.5019	0				
41736	1	0	-1.0889	0				
41738	1	0	-0.4878	0				
41739	1	0	-0.7894	0				
41742	1	0	-1.0970	0				
41743	1	0	-0.5556	0				
41748	1	0	-0.3801	0				
41873	1	0	-0.2849	0				
41874	1	0	-0.3974	0				
41876	1	0	-0.8530	0				
41877	1	0	-0.6250	0				
41878	1	0	0.0501	0				
41880	1	0	-0.9480	0				
41882	1	0	-0.2294	0				
41859	1	0	-0.0878	0				
41862	1	0	-0.0322	0				
41860	1	0	-0.4093	0				
41867	1	0	-0.8138	0				
41864	1	0	-0.8452	0				
41866	1	0	-0.4084	0				
41868	1	0	-0.8380	0				
41926	1	0	-0.3630	0				
41922	1	0	-0.7034	0				
41924	1	0	-1.3621	0				
41927	1	0	-0.6047	0				
41928	1	0	-0.8154	0				
41930	1	0	-1.0124	0				
41934	1	0	-0.7594	0				
41931	1	0	-0.2644	0				
41935	1	0	-0.8258	0				
41938	1	0	0.1874	0				
41940	1	0	-0.6596	0				
41939	1	0	-0.4495	0				
41916	4	1	0.0776	0	1.3914	0.4280	-0.5859	-1.2335
41942	4	1	0.1502	0	1.4156	0.4183	-0.5292	-1.3047

Table A-13. 2007-08 MT CRT: Item Parameter Files: Grade 8 Reading

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
42137	1	0	0.1578	0				
42142	1	0	-0.6384	0				
42139	1	0	-0.4237	0				
42144	1	0	-0.3580	0				
42141	1	0	-0.5235	0				
42140	1	0	-0.9228	0				
42143	1	0	-0.2159	0				
42163	1	0	-0.4508	0				
42166	1	0	-0.1624	0				
42165	1	0	-0.0217	0				
42168	1	0	-0.0174	0				
42170	1	0	-0.4127	0				
42171	1	0	-0.8333	0				
42175	1	0	-0.4545	0				
42174	1	0	-0.1626	0				
42178	1	0	0.1709	0				
42167	1	0	0.2638	0				
42176	1	0	-0.1253	0				
42181	1	0	-0.3386	0				
41806	1	0	-0.8330	0				
41808	1	0	-0.0754	0				
41809	1	0	-0.9844	0				
41810	1	0	-0.6434	0				
41811	1	0	-0.0926	0				
41813	1	0	-0.4966	0				
41814	1	0	0.1337	0				
42116	1	0	-0.1446	0				
42119	1	0	-0.4145	0				
42124	1	0	-0.4348	0				
42126	1	0	0.0170	0				
42127	1	0	-0.8314	0				
42125	1	0	-0.9354	0				
42113	1	0	-0.2650	0				
42018	1	0	-0.4776	0				
42021	1	0	0.2438	0				
42028	1	0	-0.3336	0				
42026	1	0	-0.4282	0				
42029	1	0	-0.7369	0				
42034	1	0	-0.4174	0				
42032	1	0	-0.8040	0				
42057	1	0	-0.8392	0				
42061	1	0	-0.5487	0				
42063	1	0	-0.4792	0				
42067	1	0	-0.0568	0				
42073	1	0	-0.8372	0				
42068	1	0	-1.0260	0				
42069	1	0	-0.5524	0				
42075	1	0	-0.2295	0				
42078	1	0	-0.1842	0				
42071	1	0	-0.4856	0				
42077	1	0	-0.1790	0				
42072	1	0	0.0410	0				
42184	4	1	0.2100	0	1.3014	0.4200	-0.4935	-1.2280
42081	4	1	0.1657	0	1.6407	0.3736	-0.6984	-1.3159

Table A-14. 2007-08 MT CRT: Item Parameter Files: Grade 10 Reading

IREF	MAX	A	B	C	D1	D2	D3	D4
42453	1	0	-0.6568	0				
42462	1	0	0.2525	0				
42459	1	0	-0.3455	0				
42464	1	0	0.4441	0				
44353	1	0	0.2358	0				
42472	1	0	-0.1873	0				
42466	1	0	-0.5114	0				
248733	1	0	-1.4594	0				
248738	1	0	-0.3560	0				
235820	1	0	-0.1793	0				
235822	1	0	0.1365	0				
249082	1	0	-0.4422	0				
248739	1	0	-0.8180	0				
235826	1	0	-0.7089	0				
235835	1	0	-0.2909	0				
248743	1	0	-0.8552	0				
248742	1	0	-1.0929	0				
249042	1	0	0.0987	0				
248752	1	0	-0.1606	0				
42415	1	0	-0.5124	0				
42419	1	0	-0.4767	0				
42420	1	0	0.0929	0				
42425	1	0	-0.1091	0				
42411	1	0	-0.4534	0				
42430	1	0	-0.1698	0				
42413	1	0	-0.2451	0				
235588	1	0	-1.1891	0				
235590	1	0	0.0509	0				
235592	1	0	-0.3769	0				
235593	1	0	-0.8871	0				
235594	1	0	0.2075	0				
235595	1	0	0.1365	0				
235596	1	0	-0.2774	0				
42545	1	0	-0.7292	0				
42561	1	0	-0.5723	0				
42558	1	0	-0.2025	0				
42560	1	0	-0.0193	0				
42563	1	0	0.0612	0				
42554	1	0	-0.9534	0				
42552	1	0	0.2309	0				
42707	1	0	0.0407	0				
42717	1	0	0.2973	0				
42710	1	0	0.2683	0				
42721	1	0	-0.0973	0				
42725	1	0	0.2175	0				
42731	1	0	-0.5361	0				
42728	1	0	-0.3038	0				
42733	1	0	-0.4952	0				
42737	1	0	-0.3089	0				
42736	1	0	-0.3137	0				
42742	1	0	0.2322	0				
42744	1	0	0.1706	0				
42453	1	0	-0.6568	0				
42462	1	0	0.2525	0				
42459	1	0	-0.3455	0				
42464	1	0	0.4441	0				
44353	1	0	0.2358	0				
248759	4	1	-0.0084	0	1.3617	0.2284	-0.5502	-1.0399
42746	4	1	0.54674	0	1.0412	0.4752	-0.5373	-0.9791

Table A-15. 2007-08 MT CRT: Item Parameter Files: Grade 4 Science

IREF	MAX	A	B	C	D1	D2	D3	D4
39247	1	0	-0.8052	0				
47553	1	0	-0.6427	0				
38541	1	0	-0.2105	0				
42802	1	0	-1.6774	0				
39318	1	0	-1.7736	0				
38546	1	0	-0.9384	0				
39196	1	0	-0.3749	0				
39180	1	0	-0.1844	0				
38579	1	0	-0.0297	0				
39193	1	0	-0.6691	0				
39121	1	0	-0.3841	0				
39275	1	0	-0.7641	0				
39060	1	0	-1.4330	0				
39285	1	0	-1.1969	0				
39149	1	0	-0.0715	0				
39307	1	0	0.1746	0				
38585	1	0	0.3836	0				
47556	1	0	-1.7810	0				
39230	1	0	-0.8251	0				
39063	1	0	-1.3034	0				
39054	1	0	-0.4017	0				
42800	1	0	0.6941	0				
38582	1	0	-0.6552	0				
39309	1	0	-0.1747	0				
39342	1	0	-0.2619	0				
39329	1	0	-0.2461	0				
42782	1	0	-1.7418	0				
39219	1	0	-0.1606	0				
39125	1	0	-1.0968	0				
39133	1	0	-1.5749	0				
39302	1	0	-0.2113	0				
39108	1	0	-0.5033	0				
42794	1	0	-0.7737	0				
39353	1	0	-0.4241	0				
47560	1	0	-1.4222	0				
38536	1	0	-1.2111	0				
47564	1	0	-0.0103	0				
38563	1	0	-0.6651	0				
39173	1	0	-0.8869	0				
39228	1	0	0.4055	0				
39225	1	0	-0.8741	0				
39190	1	0	-1.2027	0				
39270	1	0	0.0795	0				
42792	1	0	-1.5677	0				
39116	1	0	-0.8787	0				
39233	1	0	-0.0124	0				
39259	1	0	-0.9561	0				
39279	1	0	-1.0457	0				
39312	1	0	0.1115	0				
39207	1	0	-0.9242	0				
39210	1	0	-0.8721	0				
39073	1	0	-1.3140	0				
39248	1	0	-1.3146	0				
39145	4	1	-0.4599	0	0.3470	0.4538	-0.2116	-0.5892
39240	4	1	0.0448	0	0.9321	-0.0685	-0.2349	-0.6287

Table A-16. 2007-08 MT CRT: Item Parameter Files: Grade 8 Science

IREF	*MAX*	*A*	*B*	*C*	*D1*	*D2*	*D3*	*D4*
39789	1	0	-1.6314	39789				
75239	1	0	-0.7833	75239				
39818	1	0	-1.1899	39818				
39460	1	0	-0.5036	39460				
39833	1	0	0.2384	39833				
39733	1	0	-0.8867	39733				
75240	1	0	0.1841	75240				
75242	1	0	-0.9218	75242				
39551	1	0	-0.2084	39551				
39721	1	0	0.1590	39721				
39619	1	0	-1.4684	39619				
39483	1	0	-0.2667	39483				
39838	1	0	-0.6448	39838				
39577	1	0	-0.4599	39577				
39805	1	0	-0.0607	39805				
39742	1	0	0.6368	39742				
38603	1	0	-0.9403	38603				
39682	1	0	-0.4354	39682				
39501	1	0	-1.0447	39501				
39538	1	0	-0.5401	39538				
39856	1	0	0.6253	39856				
39757	1	0	-0.6420	39757				
39516	1	0	0.1744	39516				
39716	1	0	-0.9100	39716				
39540	1	0	-1.1068	39540				
39814	1	0	-0.2899	39814				
39809	1	0	-1.0781	39809				
39266	1	0	-0.4306	39266				
39562	1	0	0.0836	39562				
39610	1	0	-0.3968	39610				
39956	1	0	-0.6970	39956				
39487	1	0	0.0604	39487				
39634	1	0	-1.1786	39634				
38597	1	0	-0.7314	38597				
39771	1	0	-0.7447	39771				
39707	1	0	-1.3315	39707				
39613	1	0	-0.3050	39613				
39803	1	0	-0.6812	39803				
39528	1	0	-0.5967	39528				
39519	1	0	-0.9856	39519				
38598	1	0	-0.5623	38598				
39868	1	0	-0.1362	39868				
39824	1	0	-0.5068	39824				
39704	1	0	0.0911	39704				
39812	1	0	0.4873	39812				
38593	1	0	-0.6863	38593				
39899	1	0	-0.1968	39899				
39964	1	0	0.0481	39964				
38595	1	0	0.0398	38595				
39471	1	0	-0.2617	39471				
39783	1	0	-0.3861	39783				
39778	1	0	-0.7064	39778				
39849	1	0	-0.0856	39849				
39789	1	0	-1.6314	39789				
75239	1	0	-0.7833	75239				
39818	1	0	-1.1899	39818				
39460	1	0	-0.5036	39460				
39901	4	1	0.4629	0	0.9593	0.0860	-0.7154	-0.3299
39768	4	1	0.2123	0	0.6977	0.3134	-0.3971	-0.6140

Table A-17. 2007-08 MT CRT: Item Parameter Files: Grade 10 Science

IREF	MAX	A	B	C	D1	D2	D3	D4
40089	1	0	-1.4038	0				
40215	1	0	-0.8624	0				
38617	1	0	-0.4804	0				
40335	1	0	-0.4370	0				
40081	1	0	-0.7229	0				
40169	1	0	0.0339	0				
40344	1	0	-0.3126	0				
38621	1	0	-0.1242	0				
40294	1	0	-0.2219	0				
47595	1	0	-0.2232	0				
40137	1	0	0.5447	0				
40102	1	0	0.2049	0				
40358	1	0	-0.0431	0				
40312	1	0	0.0578	0				
40128	1	0	-0.1683	0				
40061	1	0	-0.4503	0				
40309	1	0	0.0331	0				
40212	1	0	-0.5118	0				
40047	1	0	-0.7560	0				
40401	1	0	-1.1305	0				
40181	1	0	-0.2438	0				
40290	1	0	-0.3226	0				
47588	1	0	-0.8853	0				
38615	1	0	-0.1586	0				
47587	1	0	-0.1491	0				
40409	1	0	-0.8333	0				
40131	1	0	0.0523	0				
40353	1	0	-0.7095	0				
40270	1	0	-0.2270	0				
40340	1	0	0.2593	0				
40099	1	0	0.8305	0				
39604	1	0	-0.2090	0				
40314	1	0	0.0168	0				
40050	1	0	-0.2139	0				
40205	1	0	-0.2121	0				
40096	1	0	-0.8695	0				
38607	1	0	-0.8947	0				
40149	1	0	-0.2907	0				
40406	1	0	-0.1232	0				
40323	1	0	-0.7446	0				
40292	1	0	-0.0982	0				
40113	1	0	-0.2493	0				
40348	1	0	-0.0795	0				
38619	1	0	-0.2860	0				
40277	1	0	0.6599	0				
40040	1	0	0.6082	0				
40176	1	0	-0.2547	0				
40110	1	0	0.2394	0				
47594	1	0	0.0757	0				
75716	1	0	0.4636	0				
40028	1	0	-0.2842	0				
40140	1	0	0.2373	0				
47580	1	0	-0.0633	0				
40089	1	0	-1.4038	0				
40215	1	0	-0.8624	0				
38617	1	0	-0.4804	0				
40335	1	0	-0.4370	0				
40195	4	1	0.3654	0	0.9179	0.2165	-0.5201	-0.6142
40332	4	1	0.6143	0	0.5281	0.0981	-0.0471	-0.5791

APPENDIX B—TECHNICAL ADVISORY COMMITTEE

Table B-1. 2007-08 MT CRT: 2007 Technical Advisory Committee (TAC) Members

First Name	*Last Name*	*Position*	*Department*	*Organization*
Art	Bangert, Ph.D.	Assistant Professor	Adult and Higher Education	Montana State University
Susan	Brookhart, Ph.D.	President		Brookhart Enterprises, LLC
Ellen	Forte, Ph.D.	President		edCount, LLC
Michael	Kozlow, Ph.D.	Program Director	Assessment Program	
Scott	Marion, Ph.D.	Vice-President		Center for Assessment
Stanley	Rabinowitz, Ph.D.	Program Director	Assessment & Standards Development Services	WestEd
Derek	Briggs, Ph.D.	Assistant Professor	School of Education	University of Colorado

APPENDIX C—SCIENCE STANDARD SETTING REPORT

2008
Montana Science Assessment

Standard-Setting Report

June 11 & 12, 2008

Helena, Montana

Standard-Setting Process

The standard-setting meeting to establish cut scores for the Montana CRT Science Assessment in grades 4, 8 and 10 was held on Wednesday, and Thursday June 11 & 12. Twenty-one panelists participated in the process (8 in grade 4, 6 in grade 8, and 7 in grade 10). A modified version of the Bookmark standard-setting method was used for setting standards; an overview of the method is described below.

This report is organized into three major sections, describing tasks completed prior to, during, and after the standard-setting meeting.

TASKS COMPLETED PRIOR TO THE STANDARD-SETTING MEETING

Creation of Performance Level Definitions (PLDs)

The PLDs presented to panelists provided the official description of the set of knowledge, skills, and abilities that students are expected to display in order to be classified into each achievement level. The descriptions are provided as Appendix B of this document.

Preparation of Materials for Panelists

The following materials were assembled for presentation to the panelists at the standard-setting meeting:

- Meeting agenda
- Confidentiality agreement
- Performance Level Definitions
- Assessment booklet
- Answer key
- Ordered item booklet
- Item map
- Rating form
- Evaluation form

Copies of the meeting agenda, Performance Level Definitions, Item Map, Rating form and evaluation are included in the appendices.

Preparation of Presentation Materials

The PowerPoint presentation used in the opening session was prepared prior to the meeting. A copy of the PowerPoint slides is included as Appendix C of this document

Preparation of Instructions for Facilitators Documents

A document was created for the group facilitator to refer to while working through the process. The facilitator's script is included as Appendix D.

Preparation of Systems and Materials for Analysis During the Meeting

The computational programming to carry out all analyses during the standard-setting meeting was completed and thoroughly tested prior to the standard-setting meeting.

Selection of Panelists

Panelists were selected prior to the standard-setting meeting by the Montana Department of Education. The goal was to recruit approximately 24 participants, representing a range of geographic areas, demographic groups, etc. The majority of the panelists were science teachers, for the general assessment, but some school administrators and special education teachers also participated. The actual number of participants was 21. A list of the panelists is included as Appendix E

TASKS COMPLETED DURING THE STANDARD-SETTING MEETING

Orientation

The standard-setting meeting began with a general orientation session. The purpose of the orientation was to provide background information, an introduction to the issues of standard setting, and a brief overview of the bookmark procedure and the activities that would occur during the standard-setting meeting.

Review of Assessment Materials

The first step after the opening session was for the panelists to take the test. The purpose of this step was to make sure the panelists were thoroughly familiar with what the assessment asks of students. Once panelists completed the test an answer key was distributed. At this point, panelists were encouraged to discuss any issues that came to mind regarding items or scoring.

Completion of Item Map

The purpose of the next step was to ensure that panelists became very familiar with the ordered item booklet and understood the relationships among the ordered items. The ordered item booklet contained one item per page, ordered from the easiest to the most difficult. The ordered item booklet was created by sorting items by their IRT-based difficulty values (*b* corresponding to $RP_{0.67}$ was used). A one-parameter logistic IRT model was used to calculate the $RP_{0.67}$ values.

The item map listed the items in the same order they were presented in the ordered item booklet and had spaces for the panelists to write in the knowledge, skills, and abilities required to answer correctly. There was also a space for the panelists to write in why they felt the current ordered item was more difficult than the previous one.

Each panelist stepped through the ordered item booklet, item by item, considering the knowledge, skills, and abilities students needed to complete each one. They recorded this information onto the item map along with reasons why an item was more difficult than the previous one. After they were finished working individually, panelists had an opportunity to discuss the item map as a group and make necessary additions or adjustments.

Review of PLDs and Definition of Borderline Students

Next, panelists reviewed the PLDs. This important step of the process was designed to ensure that panelists thoroughly understood the needed knowledge, skills, and abilities to be classified as *Novice, Nearing Proficiency, Proficient,* and *Advanced*. Panelists began individually then discussed the descriptions as a group, clarifying each level. Afterwards, panelists developed consensus definitions of borderline students, i.e., students who are "just able enough" to be categorized into an achievement level. Bulleted lists of characteristics for each level were generated based on the whole group discussion and posted in the room for reference throughout the bookmark process.

Round 1 Judgments

In the first round, panelists worked individually with the PLDs, the item map they completed earlier, and the ordered item booklet. Beginning with the first ordered item, and considering the skills and abilities needed to complete it, they asked themselves the question, "Would at least 2 out of 3 students performing at the borderline of *Nearing Proficiency* answer this question correctly?" Panelists considered each ordered item in turn, asking themselves the same question until their answer changed from "yes" (or predominantly "yes") to "no" (or predominantly "no"). A bookmark was placed there. Panelists then repeated the process for the other two cuts and used the provided rating form to record his/her ratings for each cut (see Appendix F).

Tabulation of Round 1 Results

After the Round 1 ratings were complete, Measured Progress staff calculated the average cut-points for the room based on Round 1 bookmark placements. This information was shared with the group to assist them in Round 2. The results of the panelists Round 1 ratings are outlined in Table 1.

Table 1: Round 1 Results of Montana Science Standard Setting

Grade	Achievement Level	Theta Cut*	Raw Score Min	Raw Score Max	Percent of Students
4	Novice		0	23	2.87
	Nearing Proficiency	-0.94	24	31	8.60
	Proficient	-0.55	32	46	52.44
	Advanced	0.21	47	60	36.09
8	Novice		0	29	19.41
	Nearing Proficiency	-0.40	30	38	28.17
	Proficient	0.05	39	47	37.30
	Advanced	0.51	48	60	15.12
10	Novice		0	26	28.19
	Nearing Proficiency	-0.29	27	35	28.83
	Proficient	0.12	36	49	36.48
	Advanced	0.83	50	61	6.50

*The minimum score necessary to make it into the achievement level.

Round 2 Judgments

The purpose of Round 2 was for panelists to discuss their Round 1 placements and revise their ratings, if necessary. Panelists shared their individual rationales for their bookmark placements in terms of the necessary knowledge and skills for each classification. Panelists were asked to pay particular attention to how their individual ratings compared to those of the others and get a sense for whether they were unusually stringent or lenient within the group. Room average cut-points were to be considered as well.

Although the panelists worked as a group, the facilitators made sure it was understood that they should set the bookmark according to their *individual* best judgments, and that they need *not* come to consensus. They were encouraged to listen to the points made by their colleagues but not feel compelled to change their bookmark placements.

Finally, panelists were given the opportunity to revise their Round 1 ratings on the rating form.

Tabulation of Round 2 Results

When Round 2 ratings were complete, Measured Progress staff calculated the average cut-points for the room and associated *impact* data. Impact data gave the percentage of students across the state that would fall into each achievement level category according to the Round 2 group average cut-points. This information was shared with the group to assist them in Round 3. The results of the panelists Round 2 ratings are outlined in Table 2.

Table 2: Round 2 Results of Montana Science Standard Setting

Grade	Achievement Level	Theta Cut*	Raw Score Min	Raw Score Max	Percent of Students
4	Novice		0	25	4.24
	Nearing Proficiency	-0.84	26	35	15.94
	Proficient	-0.38	36	47	48.06
	Advanced	0.27	48	60	31.76
8	Novice		0	29	19.41
	Nearing Proficiency	-0.42	30	36	21.21
	Proficient	-0.05	37	48	46.98
	Advanced	0.58	49	60	12.40
10	Novice		0	28	34.44
	Nearing Proficiency	-0.20	29	36	26.15
	Proficient	0.17	37	49	32.91
	Advanced	0.81	50	61	6.50

*The minimum score necessary to make it into the achievement level.

Round 3 Judgments

The purpose of Round 3 was to give panelists a final opportunity to discuss and, if necessary, modify their bookmark placements. Panelists were asked to consider all Round 2 results and the input of their colleagues. Once again, facilitators made sure panelists understood they were providing individual bookmark placements and not coming to consensus.

After the group discussions, panelists once again recorded bookmark placements on the rating form. The results of the panelists Round 3 ratings are outlined in Table 3.

Table 3: Round 3 Results of Montana Science Standard Setting

Grade	Achievement Level	Theta Cut*	Raw Score Min	Raw Score Max	Percent of Students
4	Novice		0	28	7.04
	Nearing Proficiency	-0.70	29	40	30.30
	Proficient	-0.14	41	51	48.32
	Advanced	0.56	52	60	14.34
8	Novice		0	25	11.56
	Nearing Proficiency	-0.57	26	36	29.06
	Proficient	-0.08	37	48	46.98
	Advanced	0.58	49	60	12.40
10	Novice		0	25	25.45
	Nearing Proficiency	-0.36	26	36	35.14
	Proficient	0.13	37	44	23.21
	Advanced	0.56	45	61	16.21

*The minimum score necessary to make it into the achievement level.

A graphical display of the results across grades is also provided in Figures 1 and 2. The percent of students in each performance level, based on the panelist recommendations is outlined in Figure 1, while the proportion of the total score that each performance level represents is outlined in Figure 2.

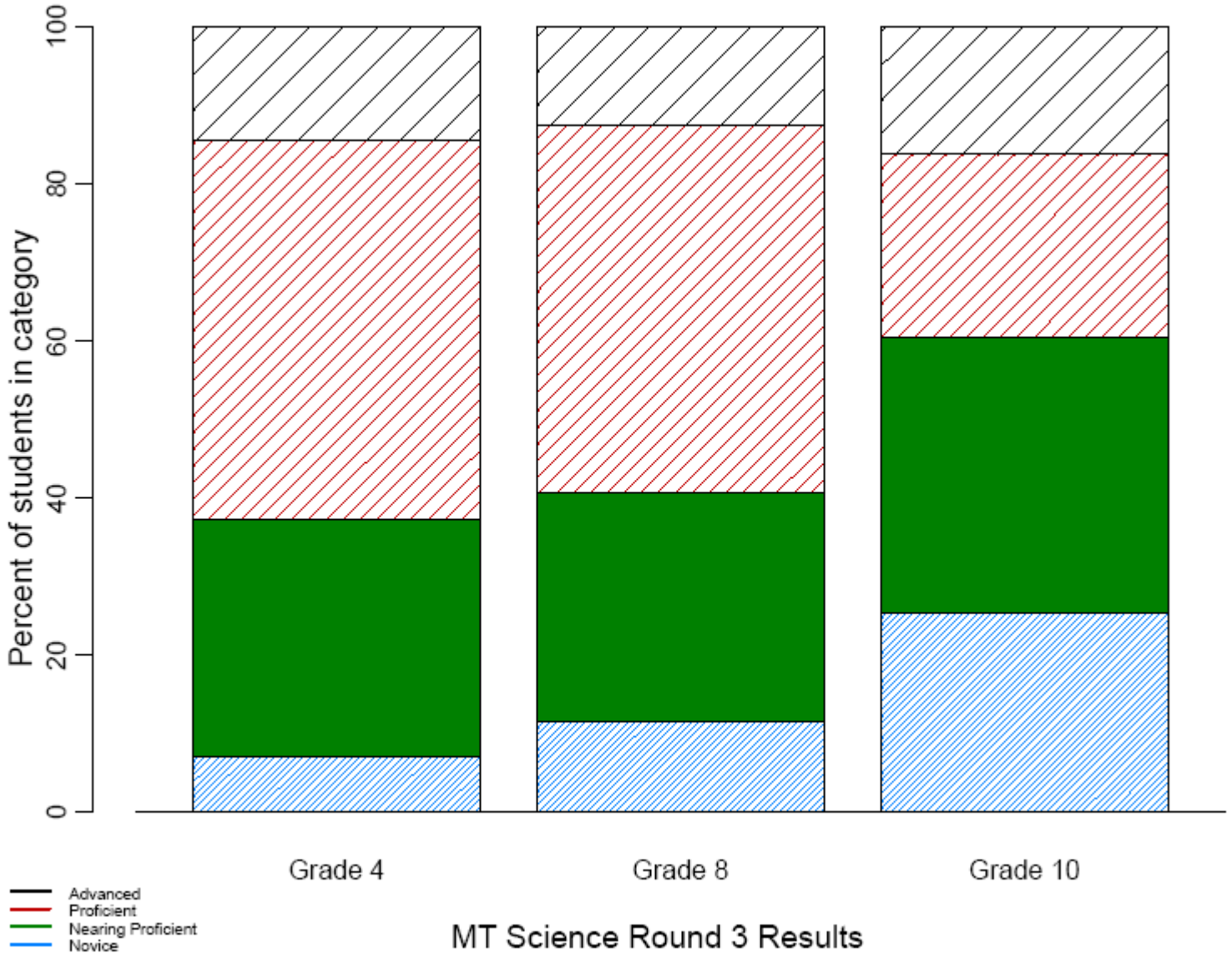


Figure 1: The percent of students falling at each performance level

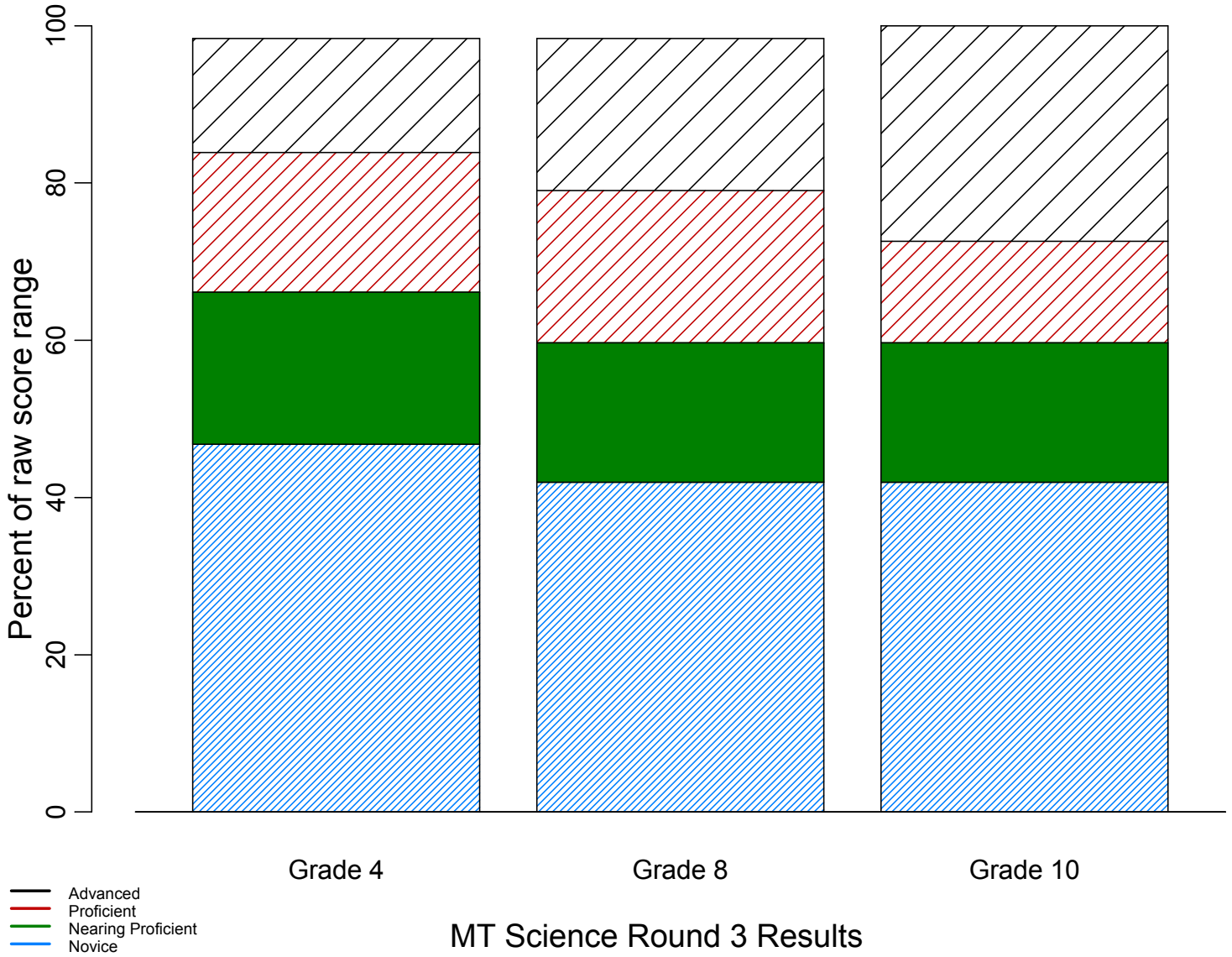


Figure 2: The percent of total raw score range for each performance level

Finally, the relationship of the panelist recommended cuts to the test is displayed in Figures 3 through 5 for grade 4, 8 and 10, respectively, using a test characteristic curve which maps the relationship between the raw score and the theta score.

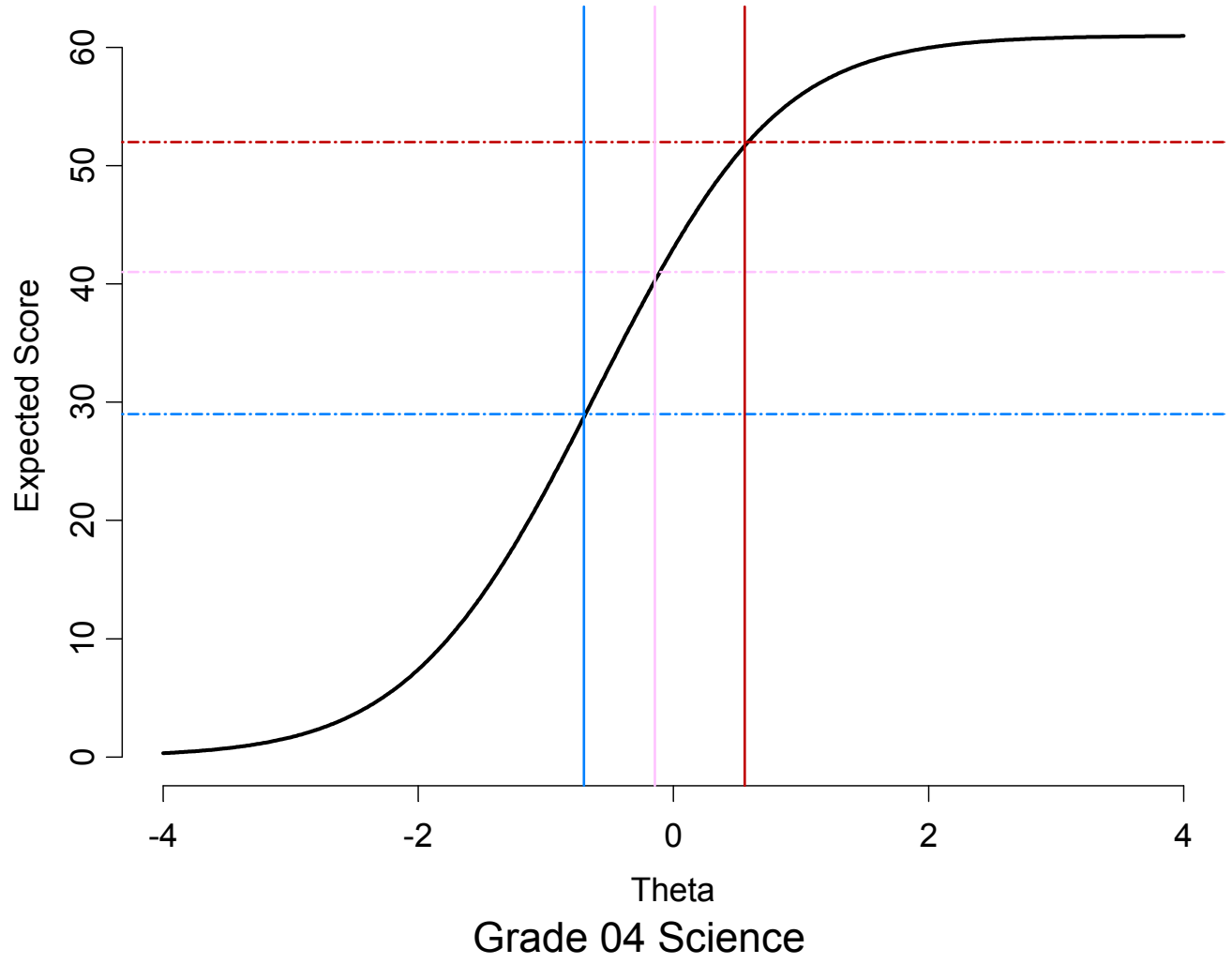


Figure 3: The relationship between the panelists recommended cuts and grade 4 test characteristic curve.

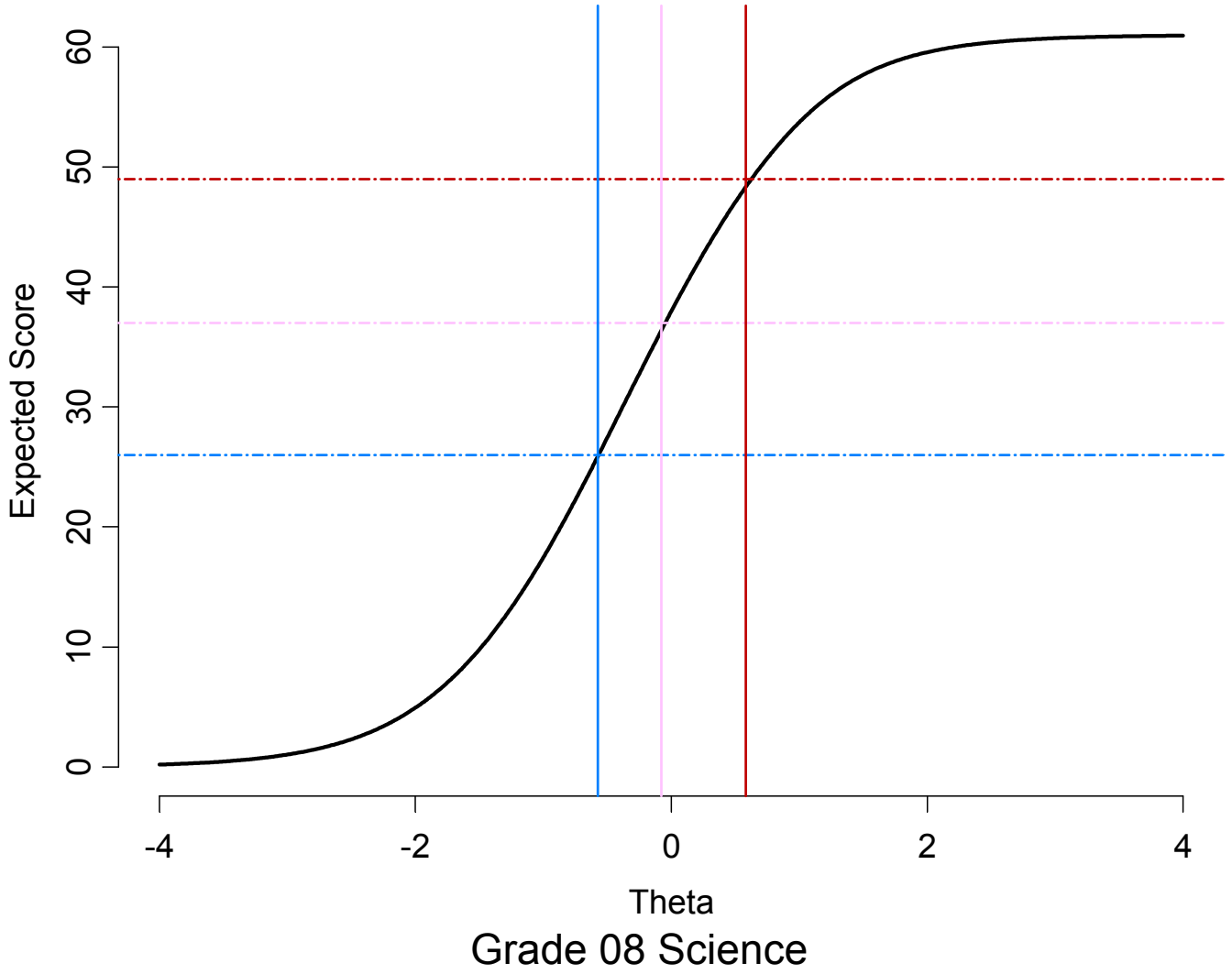


Figure 3: The relationship between the panelists recommended cuts and grade 8 test characteristic curve.

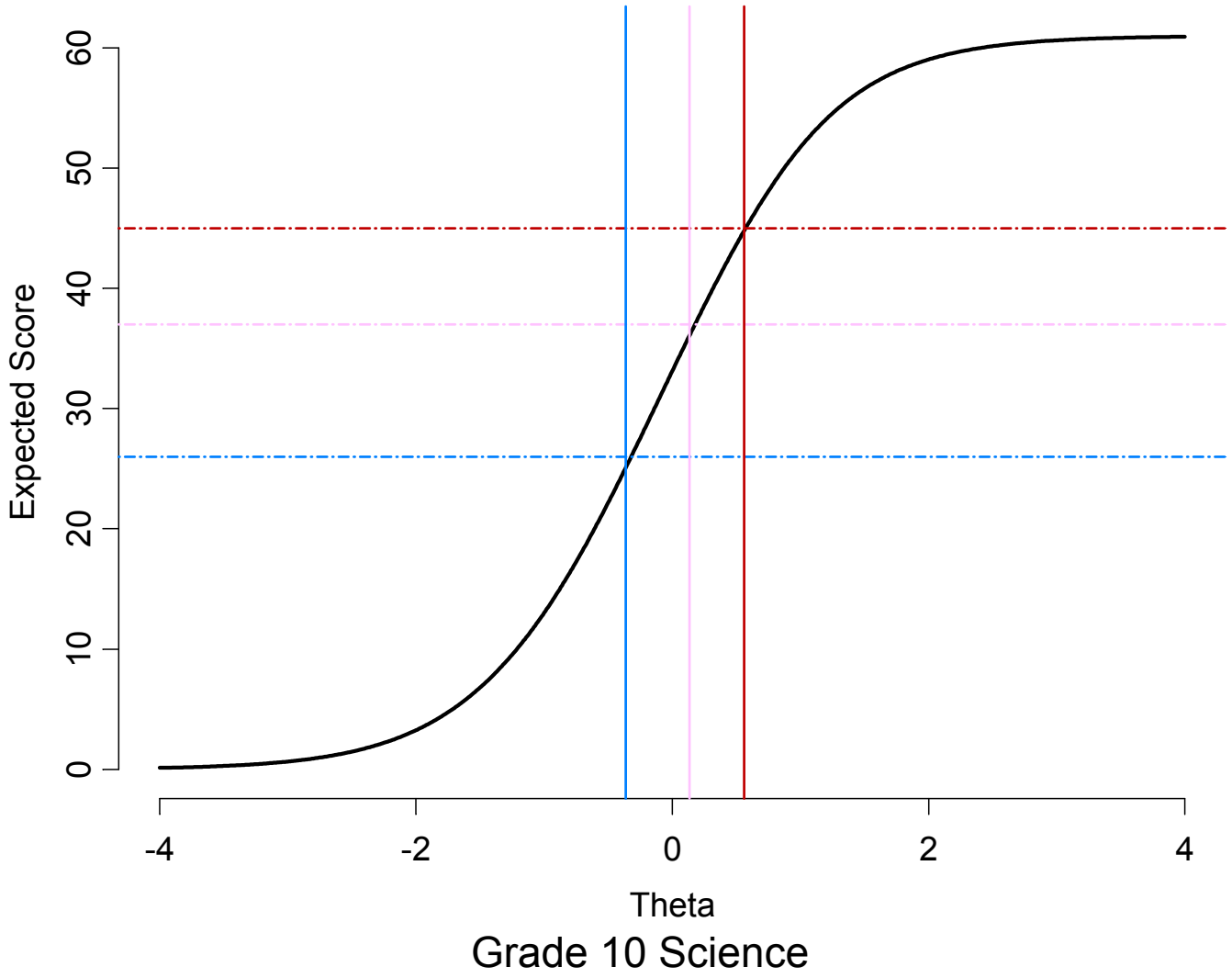


Figure 3: The relationship between the panelists recommended cuts and grade 10test characteristic curve.

Evaluation

As the last step in the standard-setting process, panelists in all three groups anonymously completed an evaluation form. A copy of the evaluation is presented as Appendix H, and the results of the evaluations are presented as Appendix I.

TASKS COMPLETED AFTER THE STANDARD-SETTING MEETING

Upon conclusion of the standard-setting meeting, several important tasks were completed. These tasks centered on reviewing the standard-setting meeting and addressing anomalies that may have occurred in the process or in the results, presenting the result to the Technical Advisory Committee (TAC), and making any final revisions or adjustments.

Analysis and Review of Panelists' Feedback

Upon completion of the evaluation forms, panelists' responses were reviewed. This review did not reveal any anomalies in the standard-setting process or indicate any reason that a particular panelist's data should not be included when the final cut-points were calculated. It appeared that all panelists understood the rating task and attended to it appropriately.

Preparation of Recommended Cut Scores

The results of the standard setting were presented to the Montana TAC meeting held on June 24th. The TAC recommended that the Round 3 results be used as the official cut points for grades four eight, and ten. Following the TAC's recommendations, OPI approved the cut points for grades four and eight. OPI requested that three sets of results for grade 10 be included in this report: the round 3 results, the round 3 results minus the standard error and the round 3 results minus twice the standard error. These results (the final results for Grades 4 and 8 and the three sets of results for Grade 10) are presented in Table 4 below.

Table 4: Results of Montana Science Standard Setting

Grade	Achievement Level	Raw Score			Percent of Students
		Theta Cut*	Min	Max	
4	Novice		0	28	7.04
	Nearing Proficiency	-0.70	29	40	30.30
	Proficient	-0.14	41	51	48.32
	Advanced	0.56	52	60	14.34
8	Novice		0	25	11.56
	Nearing Proficiency	-0.57	26	36	29.06
	Proficient	-0.08	37	48	46.98
	Advanced	0.58	49	60	12.40
10 (Round 3)	Novice		0	25	25.45
	Nearing Proficiency	-0.36	26	36	35.14
	Proficient	0.13	37	44	23.21
	Advanced	0.56	45	61	16.21
10 (Round 3 – 1SE)	Novice		0	24	22.69
	Nearing Proficiency	-0.37	25	36	37.89
	Proficient	0.13	37	44	23.21
	Advanced	0.54	45	61	16.21
10 (Round 3 – 2SE)	Novice		0	24	22.69
	Nearing Proficiency	-0.38	25	35	34.33
	Proficient	0.13	36	44	26.78
	Advanced	0.52	45	61	16.21

*The minimum score necessary to make it into the achievement level.

Preparation of Standard-Setting Report

Following final compilation of standard-setting results, Measured Progress prepared this report, which documents the procedures and results of the 2008 standard-setting meeting in order to establish performance standards for the Nevada High School Science Assessment.

APPENDIX A: AGENDA

MontCAS, Phase 2 CRT
Standard Setting Meetings

AGENDA

JUNE 11-12, 2008

WEDNESDAY, JUNE 11

8:00 – 8:30	Registration & Breakfast
8:30 – 10:30	Introduction, Overview, and Training of Standard Setting Process
10:30 – 10:45	Break
10:45 – 12:00	Move to Grade Level/Content Area Work Rooms
12:00 – 12:45	Lunch
12:45 – 2:30	Continue in Work Rooms
2:30 – 2:45	Break
2:45 – 4:00	Continue in Work Rooms
4:00	Adjourn

THRUSDAY, JUNE 12

8:00 – 8:30	Breakfast
8:30 – 10:30	Move to Grade Level/Content Area Work Rooms
10:30 – 10:45	Break
10:45 – 12:00	Continue in Work Rooms
12:00 – 12:45	Lunch
12:45 – 2:30	Continue in Work Rooms
2:30 – 2:45	Break
2:45 – 4:00	Continue in Work Rooms
4:00	Adjourn

APPENDIX B: PERFORMANCE LEVEL DESCRIPTORS

Montana K-12 Science Performance Descriptors

A Profile of Four Levels – Grade 4

The Science Performance Descriptors define students' knowledge, skills, and abilities in the science content area on a continuum from kindergarten through grade 12. These descriptions provide a picture or profile of student achievement at four performance levels: advanced, proficient, nearing proficiency, and novice.

Advanced: This level denotes superior performance.

Proficient: This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency: This level denotes that the student has partial mastery of the prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice: This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

GRADE 4 SCIENCE

Advanced: (1) A fourth-grade student at the advanced level in science demonstrates superior performance. He/she:

a. safely completes a simple investigation by asking questions, using appropriate tools and with identified variables, identifies relationships and communicates results, and identifies that observation is a key inquiry process used by Montana American Indians;

b. selects and accurately uses tools for measurement of solids, liquids, and gases, identifying properties of each state of matter and describes and models characteristics of and changes within physical and mechanical systems;

c. identifies multiple attributes of biotic (living) and abiotic (non-living) objects, including: classification based on similarities and differences; describes and models structures, functions, and processes of biotic (living) and abiotic (non-living) systems;

d. describes and explains the details of Earth's physical features and cycles;

e. discusses interactions among technology, science, and society;

f. independently identifies scientific information in the news and discusses the possible impact on local problems;

g. identifies the historical significance of scientists, discusses the impacts of their discoveries on humans today, and identifies influences of science and technology on the development of Montana American Indian cultures; and

h. identifies examples of Montana American Indian contributions to scientific and technological knowledge.

Proficient: (1) A fourth-grade student at the proficient level in science demonstrates solid academic performance. He/she:

a. with direction, safely completes a simple investigation by asking questions with identified variables, uses appropriate tools, communicates results, and identifies that observation is a key inquiry process used by Montana American Indians;

b. selects and uses tools for simple measurement of solids, liquids, and gases, identifying properties of each state of matter and describes and models characteristics of and changes within basic physical and mechanical systems;

c. identifies attributes of biotic (living) things and abiotic (non-living) objects, including: classification based on similarities and differences, basic structure and function, processes of each system;

d. Identifies and accurately illustrates Earth's features, locating several observable changes of those features;

e. identifies interactions among technology, science, and society;

f. discusses scientific information related to current events and local problems;

g. identifies the historical significance of scientists, identifies the impacts of their discoveries on humans today, and identifies influences of science and technology on the development of Montana American Indian cultures; and

h. identifies examples of Montana American Indian contributions to scientific and technological knowledge.

Nearing Proficiency: (1) A fourth-grade student at the nearing proficiency level in science demonstrates partial mastery of the prerequisite knowledge and skills fundamental for proficiency in science. He/she:

a. identifies and describes a simple investigation, and with step by step direction, given the appropriate tools, identifies and describes a simple safe investigation, and identifies that observation is a key inquiry process used by Montana American Indians;

b. with direction, effectively uses tools for simple measurement of solids, liquids, and gases, naming some properties of each state of matter and names components of basic physical and mechanical systems;

c. with direction, identifies some of biotic (living) and abiotic (non-living) objects; groups objects based on common attributes; provides basic descriptions of structure, function, and processes of a system;

d. with direction, identifies some and describes Earth's features and recognizes simple, observable changes of those features;

e. with direction, identifies some interactions among technology, science and society;

f. with direction, discusses how science plays a role in current events and local problems;

g. with direction, identifies some of the historical significance of scientists, and with direction, identifies the impacts of their discoveries on humans today, and with direction, identifies influences of science and technology on the development of Montana American Indian cultures; and

h. with direction, identifies some examples of Montana American Indian contributions to scientific and technological knowledge.

Novice: (1) A fourth-grade student at the novice level in science is beginning to attain the prerequisite knowledge and skills that are fundamental in science. He/she:

a. with direction, identifies and describes a safe, simple investigation with identified variables, and identifies that observation is a key inquiry process used by Montana American Indians;

b. with direction, identifies and uses tools for simple measurement of solids, liquids, and gases; with direction, identifies basic components of basic physical and mechanical systems;

c. with direction, identifies basic attributes of biotic (living) and abiotic (non-living) objects; groups objects based on common attributes;

d. with direction, identifies basic Earth's features and identifies fundamental changes of those features;

e. with direction, identifies how basic scientific inquiry can blend current events and local issues;

f. with direction, identifies how science plays a role in current events and local problems;

g. with direction, identifies the basic historical significance of a prominent scientist, with direction, identifies the impact of his or her discoveries on humans today, and with direction, identifies influences of science and technology on the development of Montana American Indian cultures; and

h. with direction, identifies an example of Montana American Indian contributions to scientific and technological knowledge.

Montana K-12 Science Performance Descriptors

A Profile of Four Levels – Grade 8

The Science Performance Descriptors define students' knowledge, skills, and abilities in the science content area on a continuum from kindergarten through grade 12. These descriptions provide a picture or

profile of student achievement at four performance levels: advanced, proficient, nearing proficiency, and novice.

Advanced: This level denotes superior performance.

Proficient: This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency: This level denotes that the student has partial mastery of the prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice: This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

GRADE 8 SCIENCE

Advanced: (1) An eighth-grade student at the advanced level in science demonstrates superior performance. He/she:

a. generates testable questions, safely constructs a plan for a controlled investigation, makes logical inferences based on observations, accurately interprets data by identifying the strengths and weaknesses in an investigation design, communicates results, and communicates that observation is a key inquiry process used by Montana American Indians;

b. uses physical, mental, theoretical, and mathematical models to investigate individually generated problems and/or questions about physical and chemical phenomena;

c. organizes, classifies, and describes interactions of the biotic (living) and abiotic (non-living) parts of the biosphere as well as the natural history of interactions of life on Earth and uses these skills to solve related novel (to the student) problems;

d. describes, explains and models the processes that occur in the lithosphere, hydrosphere, and atmosphere of the Earth and the universe;

e. analyzes and communicates connections and interactions among technology, science, and society by applying scientific inquiry;

f. makes informed decisions about scientific and social issues based on observations, data, analysis, and knowledge of the natural world, and effectively communicates those decisions to others;

g. independently identifies and describes examples of how science and technology are the results of human activity throughout history, independently seeks new information that connects past to present, and describes influences of science and technology on Montana American Indian cultures; and

h. describes and explains multiple examples of Montana American Indian contributions to scientific and technological knowledge.

Proficient: (1) An eighth-grade student at the proficient level in science demonstrates solid academic performance. He/she:

a. identifies and communicates testable questions, safely plans and conducts experimental investigations, communicates results, and communicates that observation is a key inquiry process used by Montana American Indians;

b. given supporting detail, describes the physical world through the application of simple chemical reactions, chemical formulas, physical, theoretical and mathematical models;

c. identifies and classifies biotic (living) things and abiotic (non-living) objects through the application of common classification schemes; identifies the interdependence of life and the environment, and explains how characteristics of living things change because of the environment;

d. describes and explains the structure and function of the Earth's lithosphere, hydrosphere, and atmosphere and the universe;

e. describes connections and interactions among technology, science, and society by applying scientific inquiry;

f. describes scientific information related to current events, and the impact on local problems;

g. independently identifies and describes examples of how science and technology are the results of human activity throughout history, seeks new information that connects past to present, and describes influences of science and technology on Montana American Indian cultures; and

h. describes and explains multiple examples of Montana American Indian contributions to scientific and technological knowledge.

Nearing Proficiency: (1) An eighth-grade student at the nearing proficiency level in science demonstrates partial mastery of the prerequisite knowledge and skills fundamental for proficiency in science. He/she:

a. with step by step direction identifies and communicates testable questions, safely plans a controlled investigation, making simple inferences based on observations and interpretation of data, and communicates that observation is a key inquiry process used by Montana American Indians;

b. gives explanations describing the physical world; through the use of simple chemical reactions, chemical formulas and physical laws, and physical models;

c. describes interactions of the biotic (living) and abiotic (non-living) parts of the biosphere; uses common classification schemes, lists examples of the interdependence of life and the environment;

d. describes the basic structure and function of the Earth's lithosphere, hydrosphere, and atmosphere and the universe;

e. with direction, describes connections and interactions among technology, science, and society by applying scientific inquiry;

f. expresses how current events impact local problems and with prompting, can discuss scientific information that effects these problems;

g. with direction, identifies and describes examples of how science and technology are the results of human activity throughout history, with direction, seeks new information that connects past to present, and describes influences of science and technology on Montana American Indian cultures; and

h. with direction, describes examples of Montana American Indian contributions to scientific and technological knowledge.

Novice: (1) An eighth-grade student at the novice level in science is beginning to attain the prerequisite knowledge and skills that are fundamental in science. He/she:

a. identifies and describes a testable question, plans for a safely controlled investigation, makes simple observations, and communicates that observation is a key inquiry process used by Montana American Indians;

b. with direction describes the physical world; identifies simple chemical reactions, chemical formulas, and demonstrates a limited understanding of physical models;
c. with direction, describes some basic interactions of the biotic (living) and abiotic (non-living) parts of the biosphere; with direction provides basic descriptions of structure and function;

d. with direction, identifies and describes the basic structure and function of the Earth's lithosphere, hydrosphere, and atmosphere and the universe;

e. with direction, identifies connections and interactions among technology, science, and society;

f. with direct instruction, can discuss basic scientific information in current events and how it impacts local problems;

g. with direction, identifies and describes examples of how science and technology are the results of human activity throughout history, and with direction, describes influences of science and technology on Montana American Indian cultures; and

h. with direction, describes examples of Montana American Indian contributions to scientific and technological knowledge.

Montana K-12 Science Performance Descriptors

A Profile of Four Levels – Grade 10

The Science Performance Descriptors define students' knowledge, skills, and abilities in the science content area on a continuum from kindergarten through grade 12. These descriptions provide a picture or

profile of student achievement at four performance levels: advanced, proficient, nearing proficiency, and novice.

Advanced: This level denotes superior performance.

Proficient: This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency: This level denotes that the student has partial mastery of the prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice: This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

GRADE 10 SCIENCE

Advanced: (1) A graduating student at the advanced level in science demonstrates superior performance. He/she:

a. formulates testable questions, safely constructs a plan, makes logical inferences, interprets data by identifying the strengths and weaknesses, communicates results, presents another investigation that more accurately assesses the topic of study, and explains that observation is a key inquiry process used by Montana American Indians;

b. creates and uses physical, mental, theoretical, and mathematical models to investigate individually generated problems and/or questions about physical and chemical phenomena;

c. creates and uses physical, mental, theoretical, and mathematical models to investigate individually generated problems and/or questions about the biotic (living) and abiotic (non-living) parts of the biosphere as well as the natural history of interactions of life on Earth and uses these skills to solve related novel (to the student) problems;

d. creates and uses physical, mental, theoretical, and mathematical models to investigate individually generated problems and/or questions about the processes that occur in the lithosphere, hydrosphere, and atmosphere of the Earth and the universe;

e. analyzes and evaluates connections and interactions among technology, science, and society by applying scientific inquiry;

f. discriminately compares scientific and social issues based on observations, data, analysis, and knowledge of the natural world, and effectively communicates those decisions to others;

g. identifies the positive and negative impacts of past, present, and future technological and scientific advances, gives possible solutions that may minimize the negative impacts on the global community, and describes and explains how science and technology apply to contemporary Montana American Indian communities; and

h. analyzes and explains Montana American Indian contributions to scientific and technological knowledge and analyzes and explains the historical impact of scientific and technological advances, including Montana American Indian examples.

Proficient: (1) A graduating student at the proficient level in science demonstrates solid academic performance. He/she:

a. generates testable questions, safely constructs a plan for a controlled investigation, makes logical inferences based on observations, accurately interprets data by identifying the strengths and weaknesses in an investigation design, communicates results, and describes and explains that observation is a key inquiry process used by Montana American Indians;

b. uses physical, mental, theoretical, and mathematical models to investigate individually generated problems and/or questions about physical and chemical phenomena;

c. organizes, classifies, and describes interactions of the biotic (living) and abiotic (non-living) parts of the biosphere as well as the natural history of

d. interactions of life on Earth and uses these skills to solve related novel (to the student) problems;

e. describes, explains and models the processes that occur in the lithosphere, hydrosphere, and atmosphere of the Earth and the universe;

f. analyzes and communicates connections and interactions among technology, science, and society by applying scientific inquiry;

g. identifies the positive and negative impacts of past, present, and future technological and scientific advances, with direction, gives possible solutions that may minimize the negative impacts on the global community, and describes and explains how science and technology apply to contemporary Montana American Indian communities; and

h. analyzes and explains Montana American Indian contributions to scientific and technological knowledge and analyzes and explains the historical impact of scientific and technological advances, including Montana American Indian examples.

Nearing Proficiency: (1) A graduating student at the nearing proficiency level in science demonstrates partial mastery of the prerequisite knowledge and skills fundamental for proficiency in science. He/she:

a. with step by step direction, safely conducts and communicates the results from simple investigations, sometimes inferring real world applications and explains that observation is a key inquiry process used by Montana American Indians;

b. identifies and constructs physical, mental, and mathematical models depicting the properties of matter in the physical world to investigate teacher-guided problems and/or questions about scientific phenomena;

c. uses models to investigate problems and/or questions about the biotic (living) and abiotic (non-living) parts of the biosphere as well as the natural history of the interactions of life on Earth;

d. with direction, describes, explains, and models the processes that occur in the lithosphere, hydrosphere, and atmosphere of the Earth and the universe;

e. identifies and describes connections and interactions among technology, science, and society by applying scientific inquiry;

f. using scientific inquiry, partially communicates interactions of science, technology, and society;

g. identifies the positive and negative impacts of past, present, and future technological and scientific advances and describes how science and technology apply to contemporary Montana American Indian communities; and

h. explains Montana American Indian contributions to scientific and technological knowledge and explains the historical impact of scientific and technological advances, including Montana American Indian examples.

Novice: (1) A graduating student at the novice level in science is beginning to attain the prerequisite knowledge and skills that are fundamental in science. He/she:

a. identifies, describes, and safely conducts a simple investigation, identifies a variable and makes real world applications, and with direction, explains that observation is a key inquiry process used by Montana American Indians;

b. with direction, identifies and uses models depicting the properties of matter in the physical world;

c. with direction, uses physical models to investigate problems and/or questions about the biotic (living) and abiotic (non-living) parts of the biosphere; describes some factors which may cause the extinction of a species;

d. with direction, describes and explains processes that occur in the lithosphere, hydrosphere, and atmosphere of the Earth and the universe;

e. identifies connections and interactions among technology, science, and society by applying scientific inquiry;

f. identifies and, with direction, communicates interactions of science, technology, and their effect on society;

g. with direction, identifies the positive and negative impacts of past, present, and future technological and scientific advances, and with direction, describes how science and technology apply to contemporary Montana American Indian communities; and

h. with direction, explains Montana American Indian contributions to scientific and technological knowledge, and with direction, describes the historical impact of scientific and technological advances, including Montana American Indian examples.

APPENDIX C: OPENING SESSION POWERPOINT

Standard Setting for MontCAS, CRT: Science, Grades 4, 8, & 10

June 11 & 12, 2008

Measured Progress

Why Are We Here?

- To recommend cut scores that distinguish between Montana's four performance levels
  - *Advanced*
  - *Proficient* ← Cut Score
  - *Nearing Proficiency* ← Cut Score
  - *Novice* ← Cut Score

Measured Progress

What is Standard Setting?

- These decisions do not come easily, that's why you are here! This process rests on your shoulders.
- We are trying to answer the questions:
  - What must a student demonstrate to be classified as Nearing Proficiency?
  - What must a student demonstrate to be classified as Proficient?
  - What must a student demonstrate to be classified as Advanced?
- Your facilitator will take you through activities that help you make informed answers.

How Will Your Answers Be Used?

- As a committee, this process will enable you to make "cut score" recommendations
- Your recommendations, along with possible adjustments, will be presented to the Montana State Board of Education for Policy Adoption.

Note

- This session is intended to be an overview
- Your facilitator will give you more details and will guide you through the process step by step

Many Standard Setting Methods

- Angoff
- Body of Work
- Bookmark

Choice of Method is Based on Many Factors

- Prior usage/history
- Recommendation/requirement by some policy making authority
- Type of assessment

- Weighing all these factors, it was determined that the Bookmark Method would be used for the Science CRT.

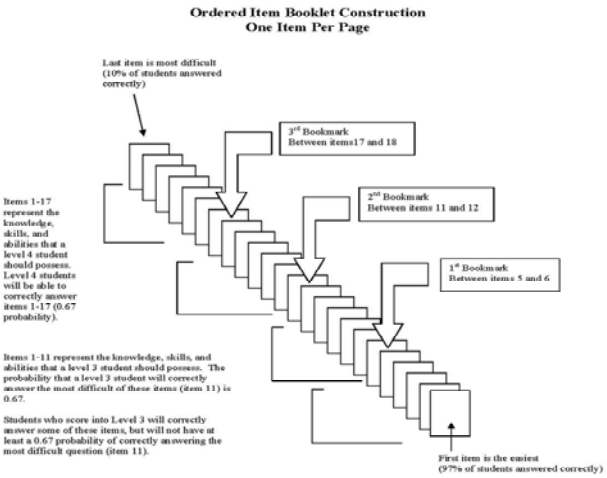
The Bookmark Procedure

- Well established procedure that has been successfully used on many assessments
- Has produced defensible results
- Appropriate for assessments that consist primarily or entirely of multiple-choice items
- Used for MT Reading and Mathematics CRT

Details for Standard Setting using the Bookmark Procedure

What is the Bookmark Procedure?

- A standard setting procedure that uses a book of items (ordered from easiest to hardest)
- Panelists place bookmarks in that book of items



How to Place a Bookmark

- A few concepts you will need to know:
  - The performance level descriptors
  - 'Borderline' students
  - The knowledge, skills, and abilities (KSAs) needed to answer each test question

How to Place a Bookmark

- Start at the beginning of the ordered item book.
- Evaluate whether at least 2 out of 3 students demonstrating skills at the 'borderline' of *Nearing Proficiency* would correctly answer item 1.
- Moving through the book, make this evaluation of each item.
- The bookmark should go where you no longer think 2 out of 3 *Nearing Proficiency* 'borderline' students would correctly answer the question.

Measured Progress

How to Place a Bookmark

Item Number	Would at least 2 out of 3 students who demonstrate skills at the *Nearing Proficiency* 'borderline' correctly answer this question?
1	Yes
2	Yes
3	Yes
4	Yes
5	Yes
6	Yes
7	Yes
8	Yes
9	No
10	No
11	No
12	No
13	No
14	No
15	No
…	No

Measured Progress

How to Place a Bookmark

- In the example, the bookmark would go between items 8 and 9
- However, it won't be that easy; there will be gray areas
- You will have opportunities to discuss your bookmark placements and change them if desired
- Place one bookmark for each of the three cut scores

Any questions about the Bookmark Procedure?

How to Place a Bookmark

- To place your bookmarks you will need to be familiar with the performance level descriptors and the assessment items

What Next?

- After this session, you will break into groups by grade level area. You will:

1. Take the assessment to familiarize yourself with the test items;
2. Complete the Item Map, which is a document that will help you with the bookmark placement process;
3. Discuss the Performance Level Descriptors and develop definitions of "borderline" *Nearing Proficiency*, *Proficient*, and *Advanced* students;

What Next?

- You will:
  1. Do the first round of bookmark placement individually, without discussion with your colleagues
  2. Discuss the first round bookmark placements as a group then do the second round of bookmark placement
  3. Discuss the second round bookmark placements along with impact data then do the final round of ratings

Note:

- It is **never** necessary for panelists to come to consensus as to where the bookmarks should be placed
- You may change your mind as a result of the discussions, or you may not
- You should be open-minded when listening to your colleagues' rationales for their ratings
- However: we want your **individual best judgment** in each round of rating

What Next?

- As the final step, we will ask you to complete an evaluation of the standard setting process

Good Luck!

APPENDIX D: FACILITATOR SCRIPT

GENERAL INSTRUCTIONS FOR MONTCAS, PHASE 2 CRT STANDARD SETTING GROUP FACILITATOR

SCIENCE: Grades 4, 8, & 10

Prior to Round 1 Ratings

Introductions:

1. Welcome group; introduce yourself (name, affiliation, a little selected background information).
2. Have each participant introduce him/herself.

Take the Test

Overview: In order to establish an understanding of the MontCAS, Phase 2 CRT science test items and for panelists to gain an understanding of the experience of the students who take the test, each participant will take the test. Panelists may wish to discuss or take issue with the items in the test. Tell them we will gladly take their feedback to the DOE. However, this is the actual assessment that students took and it is the set of items on which we must set standards.

Activities:

1) Introduce the CRT and convey/do each of the following:
    a. Tell panelists that they are about to take the actual CRT assessment.
    b. The purpose of the exercise is to help them establish a good understanding of the test items and to gain an understanding of the experience of the students who take the assessment. Let panelists know they do not need to completely answer the constructed-response questions; they can just jot a few notes.
2) Have each panelist sign the nondisclosure agreement and hand it to you.
3) Give each panelist a test booklet.
4) Tell panelists to try to take on the perspective of a student as they complete the test.
5) When the majority of the panelists have finished, pass out answer key.

Fill Out Item Map

Overview: The primary purpose of filling out the item map is for panelists to think about and document the knowledge, skills, and abilities students need to answer each question. Panelists should have an understanding of what makes one test item harder or easier than another. The notes panelists take here will be useful in helping them place their bookmarks and in discussions during the three rounds of ratings.

Activities:

1. Pass out the following materials:
   a. Item map
   b. Ordered item book

2. Provide an overview of the task paraphrasing the following:
   a. The primary purpose of this activity is for panelists to think about what makes one question harder or easier than another. For example, it may be that the concept tested is a difficult concept, or that the concept isn't difficult but that the particular wording of the question makes it a difficult question. Similarly, the concept may be a difficult one, but the wording of the question makes it easier.
   b. Panelists should take notes about their thoughts regarding each question. These will be useful in the rating activities and later discussions.

3. Tell panelists they will work individually at first. After they have completed the item map, they will then discuss it as a group.

4. Review the ordered item book and item map with the panelists. Explain what each is, and point out the correspondence of the ordered items between the two. Explain that the items are ordered from easiest to hardest. Explain that the items are ordered from easiest to hardest, and that 4-pt CRs will appear once for each possible score point.

5. Each panelist will begin with the first ordered item and compare it to the next ordered item. What makes the second item harder than the first? Panelists should not agonize over these decisions. It may be that the second item is only slightly harder than the first.

6. Panelists should work their way through the ordered item booklet, item by item, filling in the item map.

7. Once panelists have completed the item map, they should discuss them as a group. The group does not need to discuss the item maps in detail; the purpose of this step is for the panelists to discuss any particular questions or issues that arise as they are filling in the item map.

8. Based on the group discussion, the panelists should modify their own item map (make additional notes, cross things out, etc…)

Discuss Performance Level Descriptors and Describe Characteristics of the "Borderline" Student

Overview: In order to establish an understanding of the expected performance of borderline students on the test, panelists must have a clear understanding of:

1) The definition of the four performance levels, and

2) Characteristics of students who are "just able enough" to be classified into each performance level. These students will be referred to as borderline students, since they are right on the border between performance levels.

The purpose of this activity is for the panelists to obtain an understanding of the Performance Level Descriptors with an emphasis on characteristics that describe students at the borderline -- both what these students can and cannot do.

This activity is critical since the ratings panelists will be making in Rounds 1 through 3 will be based on these understandings.

Activities:

1) Introduce the task. In this activity they will:
   a. Individually review the Performance Level Descriptors;
   b. discuss the Definitions as a group; and
   c. generate bulleted lists of the characteristics of borderline ***Nearing Proficiency, Proficient*** **and** ***Advanced*** students to post in the room.

2) Pass out the Performance Level Descriptors and have panelists individually review them. Panelists can make notes if they like.

3) After individually reviewing the Descriptors, have panelists discuss each one as a group, starting with *Nearing Proficiency,* and provide clarification. The goal here is for the panelists to have a collegial discussion in which to bring up/clarify any issues or questions, and to come to a common understanding of what it means to be in each performance level. It is not unusual for panelists to disagree with the definitions they will see; almost certainly there will be some panelists who will want to change them. However, the task at hand is for panelists to have a common understanding of what knowledge, skills, and abilities (KSAs) are described by each Performance Level Descriptor. Panelists will be given an opportunity at the end of the process to provide feedback on the definitions.

4) Once panelists have a solid understanding of the Performance Level Descriptors, have them focus their discussion on the knowledge, skills, and abilities of students who are in the *Nearing Proficiency* category, but just barely. The focus should be on those characteristics and KSAs that best describe the lowest level of performance necessary to warrant a *Nearing Proficiency* classification.

5) After discussing *Nearing Proficiency*, have the panelists discuss characteristics of the borderline *Proficient* student and then characteristics of the borderline *Advanced* student. Panelists should be made aware of the importance of the *Proficient* cut.

6) Using chart paper, generate a bulleted list of characteristics for each of the levels based on the group discussion. Post these on the wall of the room.

Overview of Round 1: The purpose of Round 1 is for panelists to determine their initial bookmark placements. For this round, panelists will work individually, without any discussion with their colleagues. Starting with the cut between *Novice* and *Nearing Proficiency*, panelists will gauge the level of difficulty of each of the items for those students who barely meet the definition of *Nearing Proficiency*. Beginning with ordered item number one, the panelists will consider each item in turn and estimate whether a borderline *Nearing Proficiency* student would answer each question correctly. More specifically panelists should answer:

- Would *at least* 2 out of 3 students performing at the borderline answer the question correctly?

In the case of constructed-response questions, panelists should ask:

- Would *at least* 2 out of 3 students performing at the borderline get this score point *or higher*?

After the panelists have placed their bookmark for the *Novice /Nearing Proficiency* cut, they will then repeat the process for the *Nearing Proficiency*/*Proficient* cut and the *Proficient*/*Advanced* cut.

Activities:

1. Panelists should have their ordered item books, item maps, and the Performance Level Descriptors. Pass out one rating form to each panelist.

2. Have panelists write round number 1 and their ID number on the rating form. The ID number is on the back of their name tags.

3. Provide an overview of Round 1, covering each of the following:
    a. The primary purpose of this activity is for the panelists to determine their initial placement of each of the bookmarks. Remind panelists that they should be thinking about two-thirds of the borderline students.

    b. The panelists will work individually in this round, reviewing each of the ordered items in turn, and making a preliminary determination about where the bookmarks should be placed.

    c. Starting with the *Novice / Nearing Proficiency* cut point, the panelists should ask themselves whether students whose performance is barely *Nearing Proficiency* have at least a two-thirds chance of correctly answering each item. Each panelist should place his/her *Novice /Nearing Proficiency* bookmark where they believe the answer of 'yes' turns to 'no.'

    d. Once the panelists have placed their bookmark for the *Novice / Nearing Proficiency* cut point, they will continue through the ordered item booklet, placing their bookmarks for the *Nearing Proficiency/ Proficient* cutpoint, and, finally, for the *Proficient / Advanced* cutpoint.

e. Each panelist needs to base his/her judgments on his/her experience with the content, understanding of students, and the definitions of the borderline students generated previously.

f. If panelists are struggling with placing a particular bookmark they should use their best judgment and move on. They will have an opportunity to discuss their ratings with their colleagues and make revisions in Rounds 2 and 3.

g. Panelists should feel free to take notes if there are particular points about where they placed their bookmarks that they think are worthy of discussion in Round 2.

4. Go over the rating form with panelists.
   a. Lead panelists through a step-by-step demonstration of how to fill in the rating form.
   b. Answer questions the panelists may have about the work in Round 1.
   c. Once everyone understands what they are to do in Round 1, tell them to begin.

5. Using the ordered item book and working individually, the panelists begin with ordered item number 1. Considering each ordered item in turn, the panelists place their bookmarks for *Novice/Nearing Proficiency*, *Nearing Proficiency/Proficient*, and finally *Proficient/Advanced*.

6. As panelists complete the task, ask them to carefully inspect their rating forms to ensure they are filled out properly.
   a. **The round and ID number must be filled in**.
   b. **The item numbers identifying each cut score must be adjacent**.
   c. Check each panelist's rating form before you allow them to leave for a short break.
   d. When all the rating forms have been collected, the group will take a break. Immediately bring the rating forms to the R&A work room for tabulation.

Tabulation of Round 1 Results

Tabulation of Round 1 results will be completed by R&A as quickly as possible after receipt of the rating forms.

Round 2

Overview of Round 2: The primary purpose of Round 2 is to ask the panelists to discuss their Round 1 placements as a whole group and to revise their ratings on the basis of that discussion. They will discuss their ratings in the context of the ratings made by other members of the group. The panelists with the highest and lowest ratings should comment on why they gave the ratings they did. The group should get a sense of how much variation there is in the ratings. Panelists should also consider the question, "How tough or easy a panelist are you?" The purpose here is to allow panelists to examine their individual expectations (in terms of their experiences) and to share these expectations and experiences in order to attain a better understanding of how their experiences impact their decision-making.

To aid with the discussion, a psychometrician will present the room average bookmark placements from Round 1 to the panelists. Once panelists have reviewed and discussed their bookmark placements, they will be given the opportunity to change or revise their Round 1 ratings.

Activities:

1. Make sure panelists have their ordered item booklets, item maps, and performance level descriptors. Pass out one rating form to each panelist.

2. Have panelists write round number 2 and their ID number on the rating form.

3. A psychometrician will present the average bookmark placement for the whole group based on the Round 1 ratings. Based on their Round 1 rating form, panelists will know where they fall relative to the group average. This information is useful so that panelists get a sense if they are more stringent or more lenient than other panelists.

4. Provide an overview of Round 2. Paraphrase the following:
    a. As in Round 1, the primary purpose is to place bookmarks where you feel the performance levels are best distinguished, considering the additional information and discussion.
    b. Each panelist needs to base his/her judgments on his/her experience with the content area, understanding of students, the definitions of the borderline students generated previously, discussions with other panelists and the knowledge, skills, and abilities required to answer each item.

5. The panelists will discuss their Round 1 ratings, beginning with the first cut point.
    a. The discussion should focus on differences in where individual panelists placed their cutpoints.

    b. Panelists should be encouraged to listen to their colleagues as well as express their own points of view.

    c. If the panelists hear a logic/rationale/argument that they did not consider and that they feel is compelling, then they may adjust their ratings to incorporate that information.

    d. On the basis of the discussions and the feedback presented, panelists should make a second round of ratings.

    e. When placing their Round 2 bookmarks, panelists should not feel compelled to change their ratings.

f. The group does not have to achieve consensus. If panelists honestly disagree, that is fine. We are trying to get the best judgment of each panelist. Panelists should not feel compelled or coerced into making a rating they disagree with.

Encourage the panelists to use the discussion and feedback to assess how stringent or lenient a judge they are. If a panelist is consistently higher or lower than the group, they may have a different understanding of the borderline student than the rest of the group, or a different understanding of the Performance Level Descriptors, or both. **It is O.K. for panelists to disagree, but that disagreement should be based on a common understanding of the Performance Level Descriptors.**

6. When the group has completed their second ratings, collect the rating forms. When you collect the rating forms **carefully inspect them** to ensure they are filled out properly.
    a. **The round number and panelist ID number must be filled in**.
    b. **The item numbers identifying each cut score must be adjacent**.
    c. When all the rating forms have been collected, the group will take a break. Immediately bring the rating forms to the R&A work room for tabulation.

Round 3

Overview of Round 3: The primary purpose of Round 3 is to give the panelists one final opportunity to discuss their bookmark placements as a whole group and to revise their ratings on the basis of that discussion. Again, they will discuss their ratings in the context of the ratings made by other members of the group.

To aid with the discussion, a psychometrician will once again provide the group average Round 2 cut-point placements, as well as impact data, showing the approximate percentage of students statewide that would be classified into each performance level category based on the room average bookmark placements from Round 2.

Once panelists have reviewed and discussed their bookmark placements, they will be given a final opportunity to change or revise their ratings.

Activities:

1. Make sure panelists have their ordered item booklets, item maps, and performance level descriptors. Pass out one rating form to each panelist.

2. Have panelists write round number 3 and their ID number on the rating form.

3. A psychometrician will present and explain the following information to the panelists:
    a. The average bookmark placement for the whole group based on the Round 2 ratings. Based on their Round 2 rating form, panelists will know where they fall relative to the group average. This information is useful so that panelists get a sense if they are more stringent or more lenient than other panelists.

    b. Impact data, showing the approximate percentage of students statewide that would be classified into each performance level category based on the room average Round 2 bookmark placements.

4. Provide an overview of Round 3. Paraphrase the following:
    a. As in Round 2, the primary purpose is to place bookmarks where you feel the performance levels are best distinguished, considering the additional information and further discussion.
    b. Each panelist needs to base his/her judgments on his/her experience with the content area, understanding of students, the definitions of the borderline students generated previously, discussions with other panelists and the knowledge, skills, and abilities required to answer each item.

5. Panelists should be given a few minutes to review the Round 2 average cut points and impact data.

6. Once they have reviewed the materials, the panelists will discuss their Round 2 ratings, beginning with the first cut point.
    a. The discussion should focus on differences in where individual panelists placed their cutpoints.

    b. Panelists should be encouraged to listen to their colleagues as well as express their own points of view.

c. If the panelists hear a logic/rationale/argument that they did not consider and that they feel is compelling, then they may adjust their ratings to incorporate that information.

d. On the basis of the discussions and the feedback presented, panelists should make a third round of ratings.

e. When placing their Round 3 bookmarks, panelists should not feel compelled to change their ratings.

f. The group does not have to achieve consensus. If panelists honestly disagree, that is fine. We are trying to get the best judgment of each panelist. Panelists should not feel compelled or coerced into making a rating they disagree with.

Encourage the panelists to use the discussion and feedback to assess how stringent or lenient a judge they are. If a panelist is consistently higher or lower than the group, they may have a different understanding of the borderline student than the rest of the group, or a different understanding of the Performance Level Descriptors, or both. **It is O.K. for panelists to disagree, but that disagreement should be based on a common understanding of the Performance Level Descriptors.**

7. When the group has completed their final ratings, collect the rating forms. When you collect the rating forms **carefully inspect them** to ensure they are filled out properly.
   a. **The round number and panelist ID number must be filled in**.
   b. **The item numbers identifying each cut score must be adjacent**.
   c. Immediately provide the completed rating forms to R&A. The panelists will not see the results from this round.

Feedback on Performance Level Descriptors

After completing the third round of ratings, panelists will be given an opportunity to provide suggested modifications or enhancements to the performance level descriptors to reflect the specific knowledge, skills, and abilities required to be classified into each level. Panelists may also recommend edits reflecting skills that were included on the assessment but did not appear in the performance level descriptors, or vice versa. Make sure panelists know that these are recommendations and that they may not all be implemented.

Complete Evaluation Form

Upon completion of Round 3, have panelists fill out the evaluation form. Emphasize that their honest feedback is important.

APPENDIX E: PANNELIST AFFILIATIONS

GRADE 4:

Kathy	Gaul
Christi	Hoskinson
Carol	Kron
Vicky	Michels
Karen	Miller
Carol	Morgan
Mavis	Peterson
Patti	Vennes

GRADE 8:

Carl	Christiansen
Michael	Howard
David	Pettit
Sue	Degooyer
Susan	Luinstra
Kris	Goyins

GRADE 10:

Steve	Bell
Holly	Faris
Robin	Hompesch
Allyson	Hoof
Don	Samuelson
Dawn	Sturman
Chris	West

APPENDIX F: SAMPLE RATING FORM

MontCAS Science Grade 4
Rating Form

Round _____

ID _____

Novice Ordered Item Numbers		**Nearing Proficiency** Ordered Item Numbers		**Proficient** Ordered Item Numbers		**Advanced** Ordered Item Numbers	
First	Last	First	Last	First	Last	First	Last
1	___	___	___	___	___	___	**60**

<u>Directions</u>: Please enter the range of ordered item numbers that fall into each performance level category according to where you placed your cutpoints.

<u>Note</u>: The ranges *<u>must</u>* be adjacent to each other. For example: Novice: 1 – 13, Nearing Proficiency: 14 – 34, Proficient: 35 – 45, Advanced: 46 – 60.

APPENDIX G: SAMPLE ITEM MAP

Grade 4 Science
Item Map

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
1	47556			1
2	39318			1
3	42782			1
4	42802			1
5	39133			1
6	42792			1
7	39060			1
8	47560			1
9	39145			1
10	39248			1
11	39073			1
12	39063			1
13	38536			1
14	39190			1
15	39285			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
16	39125			1
17	39240			1
18	39279			1
19	39145			2
20	39259			1
21	38546			1
22	39207			1
23	39173			1
24	39116			1
25	39225			1
26	39210			1
27	39230			1
28	39247			1
29	42794			1
30	39275			1
31	39193			1
32	38563			1
33	38582			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
34	47553			1
35	39108			1
36	39145			3
37	39353			1
38	39054			1
39	39121			1
40	39196			1
41	39342			1
42	39240			2
43	39329			1
44	39302			1
45	38541			1
46	39180			1
47	39309			1
48	39219			1
49	39149			1
50	38579			1
51	39233			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
52	47564			1
53	39270			1
54	39312			1
55	39240			3
56	39307			1
57	39145			4
58	38585			1
59	39228			1
60	42800			1
61	39240			4

Grade 8 Science
Item Map

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
1	39789			1
2	39619			1
3	39707			1
4	39818			1
5	39634			1
6	39540			1
7	39809			1
8	39501			1
9	39519			1
10	38603			1
11	75242			1
12	39716			1
13	39733			1
14	39768			1
15	75239			1
16	39771			1
17	38597			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
18	39778			1
19	39956			1
20	38593			1
21	39901			1
22	39803			1
23	39838			1
24	39757			1
25	39528			1
26	38598			1
27	39538			1
28	39824			1
29	39460			1
30	39577			1
31	39682			1
32	39266			1
33	39610			1
34	39783			1
35	39613			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
36	39814			1
37	39483			1
38	39471			1
39	39768			2
40	39551			1
41	39899			1
42	39868			1
43	39849			1
44	39805			1
45	38595			1
46	39964			1
47	39487			1
48	39562			1
49	39704			1
50	39721			1
51	39901			2
52	39516			1
53	75240			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
54	39833			1
55	39768			3
56	39812			1
57	39856			1
58	39742			1
59	39901			3
60	39768			4
61	39901			4

Grade 10 Science
Item Map

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
1	40089			1
2	40401			1
3	38607			1
4	47588			1
5	40096			1
6	40215			1
7	40409			1
8	40195			1
9	40047			1
10	40323			1
11	40081			1
12	40353			1
13	40212			1
14	38617			1
15	40061			1
16	40335			1
17	40290			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
18	40344			1
19	40149			1
20	38619			1
21	40028			1
22	40332			1
23	40176			1
24	40113			1
25	40181			1
26	40270			1
27	47595			1
28	40294			1
29	40050			1
30	40205			1
31	39604			1
32	40128			1
33	38615			1
34	47587			1
35	38621			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
36	40406			1
37	40292			1
38	40348			1
39	47580			1
40	40358			1
41	40195			2
42	40314			1
43	40309			1
44	40169			1
45	40131			1
46	40312			1
47	47594			1
48	40102			1
49	40332			2
50	40140			1
51	40110			1
52	40340			1
53	75716			1

Item Order	IABS	What does this item measure?	Why is this item more difficult than the preceding item?	Score Point
54	40137			1
55	40332			3
56	40195			3
57	40040			1
58	40277			1
59	40099			1
60	40195			4
61	40332			4

APPENDIX H: EVALUATION

MONTANA MONTCAS EVALUATION FORM

Standard Setting 2008

1. What is your overall impression of the process used to set performance standards for MontCAS? *(Circle one)*

   A. Very Good
   B. Good
   C. Unsure
   D. Poor
   E. Very Poor

2. How clear were you with the performance level descriptors? *(Circle one)*

   A. Very Clear
   B. Clear
   C. Somewhat Clear
   D. Not Clear

3. How would you judge the length of time of this meeting for setting performance standards? *(Circle one)*

   A. About right
   B. Too little time
   C. Too much time

4. What factors influenced the standards you set? (For each, circle the most appropriate rating from 1=Not at all Influential to 5=Very Influential)

   A. The performance level descriptors

Not at all Influential		Moderately Influential		Very Influential
1	2	3	4	5

   B. The assessment items

Not at all Influential		Moderately Influential		Very Influential
1	2	3	4	5

   C. Other panelists

Not at all Influential		Moderately Influential		Very Influential
1	2	3	4	5

   D. My experience in the field

Not at all Influential		Moderately Influential		Very Influential
1	2	3	4	5

E. Other (*please specify*_____)

Not at all Influential		Moderately Influential		Very Influential
1	2	3	4	5

5. Do you believe the cut scores set by the panel are correctly placed on the assessment score scale?

A. Definitely Yes
B. Probably Yes
C. Unsure
D. Probably No
E. Definitely No

Please explain your answer:

_____

_____

_____

For each statement below, please circle the rating that best represents your judgment

6. The opening session was:

Not at all Useful				Very Useful
1	2	3	4	5

7. The performance level descriptors were:

Not at all Clear				Very Clear
1	2	3	4	5

8. Providing additional details to the performance level descriptors was:

Not at all Useful				Very Useful
1	2	3	4	5

9. The discussion with other panelists was:

Not at all Useful				Very Useful
1	2	3	4	5

10. The standard setting task was:

Not at all Clear				Very Clear
1	2	3	4	5

11. The impact data at the beginning of round 3 was:

Not at all Useful				Very Useful
1	2	3	4	5

12. How could the standard setting process have been improved?

_____
_____
_____
_____
_____

Additional Comments

13. Please provide any additional comments or suggestions about the standard setting process.

_____
_____
_____
_____
_____
_____

APPENDIX I: EVALUATION RESULTS

GRADE 4 SCIENCE					
	Very Good	Good	Unsure	Poor	Very Poor
What is your overall impression of the process used to set performance standards for MontCAS?? *(Circle one)*	7	0	1	0	0
How clear were you with the performance level descriptors? *(Circle one)*	4	4	0	0	0
How would you judge the length of time of this meeting for setting performance standards? *(Circle one)*	8	0	0	0	0
What factors influenced the standards you set? (For each, circle the most appropriate rating from 1=Not at all Influential to 5=Very Influential)	Not at all Influential 1	2	3	4	Very Influential 5
A. The Performance Level Descriptors	0	0	0	2	6
B. The assessment items	0	0	0	3	5
C. Other panelists	0	0	2	1	5
D. My experience in the field	0	0	1	1	6
E. Other (specify) ~ where we listed why one item was more difficult than the preceding item	1	0	0	0	0
E1. Other (specify) ~ Students	0	0	0	1	0
E2. Other (specify) ~ Data	0	0	0	0	1
	Definitely Yes	Probably Yes	Unsure	Probably No	Definitely No
For this grade level do you believe the cut scores set by the panel are correctly placed on the assessment score scale?	0	7	1	0	0

Please explain your answer:
~ I think we really thought about all the factors and discussed everything in depth. I don't know if you could ever say it's definitely placed where they should be – it's not an exact thing – too many factors come into play. ~ I think there were 1 or 2 panelists not likely to change scores based on discussion. ~ We don't know precisely where the cuts came – are they correctly placed? When I see my students' results I can answer better. ~ I am having difficulty with some of the questions. Not knowing how the cut scores came out causes questions. ~ Without knowing the round 3 outcome ~ It was difficult because of the quality of some of the questions and where they were placed ~ Not knowing what round 3 scores are, I cannot say definitely. However, analyzing #1 & #2 and the group's comments help me to say probably yes.

For each statement below, please circle the rating that best represents your judgment.						
A.	The opening session was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	2	4	1
B.	The Performance Level Descriptors were:	Not at all Clear 1	2	3	4	Very Clear 5
		0	0	1	4	3
C.	Providing additional details to the Performance Level Descriptors was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	0	0	8
D.	The discussion with other panelists was:	Not at all Clear 1	2	3	4	Very Clear 5
		0	0	0	1	7

E.	The standard setting task was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	0	3	5
F.	The impact data at the beginning of round 3 was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	0	0	8

How could the standard setting process have been improved?

~ I think it is a good process – lots of discussion and years of teaching experience coming together
~ I think the book mark process works well. It seems to be more objective
~ I would like to see the results from my class while – or at some point in a round – to see if my perceptions of what is novice – NP – P – A – hold somewhat true
~ Experience in the entire process (other sessions) is very beneficial
~ Get more people involved, especially from reservation schools
~ Displayed impact data
~ Very interesting! Enjoyed learning the process
~ The novice and nearing proficiency performance level descriptors are so close that it is hard to differentiate in some areas

Please provide any additional comments or suggestions about the standard setting process?

~ Discuss is the most important part of the process. It made the setting of cut off scores from fuzzy to very clear.
~ It was very enjoyable and provided information to me. Even though science literacy is important, we need to assess the vocabulary used at the grade 4 level on several of the questions. It is a terrific process! Thank you all.
~ We had a super facilitator. She was very capable of getting opinionated teachers back on track. Her summary skills were excellent. Thanks!
~ Throw out the test and let us get back to teaching. Sorry.
~ Interesting in that this group of educators was very concerned that they balance the cuts between us and the questions
~ This is a valuable process for me as a teacher, and I wish more of my colleagues would participate

GRADE 8 SCIENCE					
	Very Good	Good	Unsure	Poor	Very Poor
What is your overall impression of the process used to set performance standards for MontCAS?? *(Circle one)*	4	1	0	0	0
How clear were you with the performance level descriptors? *(Circle one)*	3	2	0	0	0
How would you judge the length of time of this meeting for setting performance standards? *(Circle one)*	5	0	0	0	0
What factors influenced the standards you set? (For each, circle the most appropriate rating from 1=Not at all Influential to 5=Very Influential)	Not at all Influential 1	2	3	4	Very Influential 5
A. The Performance Level Descriptors	0	0	1	2	2
B. The assessment items	0	0	0	3	2
C. Other panelists	0	1	0	4	0
D. My experience in the field	0	0	0	2	1
E. Other (specify) ~ life experience/teaching experience	0	0	0	1	0
	Definitely Yes	Probably Yes	Unsure	Probably No	Definitely No
For this grade level do you believe the cut scores set by the panel are correctly placed on the assessment score scale?	1	4	0	0	0
Please explain your answer:					

~ after discussions – cut scores are correctly placed
~ Panel very articulate – lots of experience different areas of state!!
~ Interactive input for all members
~ I still feel the cut score between Nearing Proficient and Proficient should have been nearer 24

For each statement below, please circle the rating that best represents your judgment.						
A.	The opening session was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	1	1	2
B.	The Performance Level Descriptors were:	Not at all Clear 1	2	3	4	Very Clear 5
		0	0	0	2	3
C.	Providing additional details to the Performance Level Descriptors was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	0	3	2
D.	The discussion with other panelists was:	Not at all Clear 1	2	3	4	Very Clear 5
		0	1	0	0	4
E.	The standard setting task was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	0	2	3
F.	The impact data at the beginning of round 3 was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	0	1	4

How could the standard setting process have been improved?
~ Perhaps more teachers to set and get more of a variety of ideas – there were 6 of us, perhaps 8 to 10 ~ good!
Please provide any additional comments or suggestions about the standard setting process?
~ Great experience ~ This was a good experience, very enlightening and encouraging ~ Good group leader – excellent group participation

GRADE 10 SCIENCE					
	Very Good	Good	Unsure	Poor	Very Poor
What is your overall impression of the process used to set performance standards for MontCAS?? *(Circle one)*	6	1	0	0	0
How clear were you with the performance level descriptors? *(Circle one)*	5	2	0	0	0
How would you judge the length of time of this meeting for setting performance standards? *(Circle one)*	7	0	0	0	0
What factors influenced the standards you set? (For each, circle the most appropriate rating from 1=Not at all Influential to 5=Very Influential)	Not at all Influential 1	2	3	4	Very Influential 5
A. The Performance Level Descriptors	0	0	1	1	5
B. The assessment	0	0	2	1	4
C. Other panelists	0	0	1	1	5
D. My experience in the field	0	0	1	0	6
E. other (specify) ~ great facilitators	0	0	0	0	1
E. Other (specify) ~ group discussions * probably the most influential part of this process was the cross pollination between teachers in our discussion about how to determine performance standards	0	0	0	0	1
E. Other (specify) ~ expectations by NCLB/by state/by teachers ("politics")	0	0	0	0	1
E. Other (specify) ~ order of the questions	0	0	1	0	0
E. Other (specify) ~ my own children – ages 14 & 16	0	0	0	1	0

	Definitely Yes	Probably Yes	Unsure	Probably No	Definitely No
For this grade level do you believe the cut scores set by the panel are correctly placed on the assessment score scale?	0	6	1	0	0
Please explain your answer:					
~ but am confident in our cut scores as fairly representing the 10th grade students in MT ~ We weren't in 100% agreement on all cuts. The "after round 2" statistics were helpful. But we did change the cuts to give our expectations of what should be vs. reflecting what is? ~ Agreement that some questions (mitosis) were not useful for borderline but within 2-3 questions…. Concern this is 5-10% ~ Would like to know % after final round. Never 100% sure, I like what we did ~ I think there could always be improvement – but I do believe we did the best we could ~ Original cut scores overestimated, Round 3 more appropriate					
For each statement below, please circle the rating that best represents your judgment.					
A. The opening session was:	Not at all Useful 1	2	3	4	Very Useful 5
	0	1	0	1	5
B. The Performance Level Descriptors were:	Not at all Clear 1	2	3	4	Very Clear 5
	0	0	2	2	3
C. Providing additional details to the Performance Level Descriptors was:	Not at all Useful 1	2	3	4	Very Useful 5
	0	0	0	2	5
D. The discussion with other panelists was:	Not at all Clear 1	2	3	4	Very Clear 5
	0	0	0	0	7

E.	The standard setting task was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	0	1	6
F.	The impact data at the beginning of round 3 was:	Not at all Useful 1	2	3	4	Very Useful 5
		0	0	0	1	6

How could the standard setting process have been improved?

~ The discussions were most appropriate and needed, very appropriate scheduling, well done!
~ I do not know, this was an exceptional experience.
~ "bucket" info was misleading. Chloe was terrific! Leaning on without leading… circling back… getting responses from all.
~ I wish Montana simply used the National Science Standards instead of writing and maintaining our own.
~ This was well worth my time and effort. Thank you for providing the opportunity to grow and learn from this process!

Please provide any additional comments or suggestions about the standard setting process?

~ item order seemed out of place due to degree of difficulty so other factors besides just the questions were also influential)
~ Chloe was a fantastic facilitator! She gave guidance at the appropriate times. She was non-judgmental and engaging. Well done!
~ Great job!
~ Great facilitators
~ Just because MT doesn't need to re-invent the wheel and as a parent I want to see my child's progress in relation to the rest of the Nation not just the state
~ Maybe look at using school counselors to help with group definitions. Also could a student group be used?
~ Having chem. Up (Junior, Sr. teachers) was ?? teachers of Sophomores, 14-16 years… realistic day to day experiences w/ 9, 10… high exp. Of these volunteer teachers so (ADV) was limited and acceptance of more Novice – NP tan "should".
~ Our group found some of the questions on the test to have multiple correct answers. We've got to do something about MT standards at the "12th grade." Not all students take physics & chemistry. Therefore, our described KSA @ grade 12 are unrealistic.
~ Thank you.

APPENDIX D—CRT PERFORMANCE LEVEL DESCRIPTORS AND STUDENT DISTRIBUTIONS WITHIN RAW- AND SCALE-SCORE RANGES

Table D-1. 2007-08 Montana CRT: Performance Level Descriptors (General)

Advanced	This level denotes superior performance.
Proficient	This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
Nearing Proficiency	This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.
Novice	This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

Table D-2. 2007-08 Montana CRT: Student Distributions within Performance Level Raw- and Scale-Score Ranges—Grade 3

	Reading			Mathematics		
	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students
Advanced	44-60	287-300	4.0%	55-66	290-300	17.0%
Proficient	30-43	250-286	11.8%	43-54	250-289	19.5%
Nearing Proficiency	20-29	225-249	41.9%	34-42	225-249	38.3%
Novice	0-19	200-224	42.3%	0-33	200-224	25.2%

Table D-3. 2007-08 Montana CRT: Student Distributions within Performance Level Raw- and Scale-Score Ranges—Grade 4

	Reading			Mathematics			Science		
	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students
Advanced	44-60	289–300	4.5%	50-66	291–300	13.8%	52-61	281-300	7.1%
Proficient	29-43	250–288	15.9%	37-49	250–290	18.9%	41-51	250–280	30.3%
Nearing Proficiency	19-28	225–249	45.3%	28-36	225–249	40.2%	29-40	225–249	48.3%
Novice	0-18	200–224	34.3%	0-27	200–224	27.0%	0-28	200–224	14.3%

Table D-4. 2007-08 Montana CRT: Student Distributions within Performance Level Raw- and Scale-Score Ranges—Grade 5

	Reading			Mathematics		
	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students
Advanced	43-60	287–300	6.7%	49-66	289–300	13.1%
Proficient	31-42	250–286	11.2%	35-48	250–288	18.8%
Nearing Proficiency	22-30	225–249	30.8%	26-34	225–249	42.0%
Novice	0-21	200–224	51.3%	0-25	200–224	26.1%

Table D-5. 2007-08 Montana CRT: Student Distributions within Performance Level Raw- and Scale-Score Ranges—Grade 6

	Reading			Mathematics		
	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students
Advanced	44-60	289–300	5.0%	46-66	287–300	16.3%
Proficient	30-43	250–288	11.0%	32-45	250–286	20.0%
Nearing Proficiency	21-29	225–249	38.4%	24-31	225–249	38.0%
Novice	0-20	200–224	45.6%	0-23	200–224	25.7%

Table D-6. 2007-08 Montana CRT: Student Distributions within Performance Level Raw- and Scale-Score Ranges—Grade 7

	Reading			Mathematics		
	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students
Advanced	46-60	288–300	6.2%	42-66	289–300	12.8%
Proficient	31-45	250–287	10.1%	28-41	250–288	20.0%
Nearing Proficiency	22-30	225–249	38.8%	20-27	225–249	37.4%
Novice	0-21	200–224	44.9%	0-19	200–224	29.8%

Table D-7. 2007-08 Montana CRT: Student Distributions within Performance Level Raw- and Scale-Score Ranges—Grade 8

	Reading			Mathematics			Science		
	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students
Advanced	47-60	289–300	8.0%	46-66	283–300	14.1%	49-61	283–300	11.6%
Proficient	34-46	250–288	10.2%	32-45	250–282	25.9%	37-48	250–282	29.1%
Nearing Proficiency	26-33	225–249	35.9%	21-31	225–249	34.1%	26-36	225–249	47.0%
Novice	0-25	200–224	45.9%	0-20	200–224	26.0%	0-25	200–224	12.4%

Table D-8. 2007-08 Montana CRT: Student Distributions within Performance Level Raw- and Scale-Score Ranges—Grade 10

	Reading			Mathematics			Science		
	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students	Raw Score Range	Scale Score Range	Percentage of Students
Advanced	46-60	289–300	8.2%	43-66	281–300	11.4%	45-61	269–300	22.7%
Proficient	33-45	250–288	13.3%	28-42	250–280	35.4%	36-44	250–268	34.3%
Nearing Proficiency	25-32	225–249	43.2%	17-27	225–249	35.0%	25-35	225–249	26.8%
Novice	0-24	200–224	35.3%	0-16	200–224	18.1%	0-24	200–224	16.2%

APPENDIX E—REPORT SHELLS

MontCAS, Phase 2
CRT

System:
Grade: 04
Spring 2008

Mathematics

System Summary Report

I. Distribution of Scores

Perf. Level	Scores	System			State		
		Number	% of Students	% of Students in Cat.	Number	% of Students	% of Students in Cat.
Advanced	299-300						
	297-298						
	295-296						
	293-294						
	291-292						
Proficient	283-290						
	275-282						
	266-274						
	258-265						
	250-257						
Nearing Proficiency	245-249						
	240-244						
	235-239						
	230-234						
	225-229						
Novice	220-224						
	215-219						
	210-214						
	205-209						
	200-204						

II. Subtest Results

Mathematics		Possible Points	Average Points Earned	
			System	State
Total Points		66		
Standards	1. Problem Solving	This standard is assessed within the frameworks of standards 2-7.		
	2. Numbers and Operations	22		
	3. Algebra	8		
	4. Geometry	10		
	5. Measurement	10		
	6. Data Analysis, Statistics, and Probability	8		
	7. Patterns, Relations, and Functions	8		

CRT Performance Level Descriptors

Advanced (291-300)
This level denotes superior performance.

Proficient (250-290)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2 CRT

Confidential

Mathematics **System Summary Report**

System:
Grade: 04
Spring 2008

III. Results for Subgroups of Students

Reporting Category	System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students										
Gender										
Male										
Female										
Ethnicity										
American Indian or Alaska Native										
Asian										
Hispanic										
Black or African American										
Native Hawaiian or Other Pacific Islander										
White										
Special Education										
Students with a 504 Plan										
Title I (optional)										
Tested with Standard Accommodation										
Tested with Non-Standard Accommodation										
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report								
Migrant										
Gifted/Talented										
LEP/ELL										
Former LEP Student										
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students								
Free/Reduced Lunch										
Significant Cognitive Disability	Data not available for the 2008 report									
Special Education Disability(ies):										
Autism										
Cognitive Delay										
Deaf-Blindness Impairment										
Deafness										
Emotional Disturbance										
Hearing Impairment										
Learning Disability										
Other Health Impairment										
Orthopedic Impairment										
Speech/Language										
Traumatic Brain Injury										
Visual Impairment										

* Less than ten (10) students were assessed

MontCAS, Phase 2
CRT

System:
Grade: 04
Spring 2008

Reading

System Summary Report

I. Distribution of Scores

Perf. Level	Scores	System			State		
		Number	% of Students	% of Students in Cat.	Number	% of Students	% of Students in Cat.
Advanced	299-300						
	296-298						
	294-295						
	291-293						
	289-290						
Proficient	281-288						
	273-280						
	266-272						
	258-265						
	250-257						
Nearing Proficiency	245-249						
	240-244						
	235-239						
	230-234						
	225-229						
Novice	220-224						
	215-219						
	210-214						
	205-209						
	200-204						

II. Subtest Results

Reading		Possible Points	Average Points Earned	
			System	State
Total Points		60		
Standards	1. Students construct meaning as they comprehend, interpret, and respond to what they read	21		
	2. Students apply a range of skills and strategies to read	19		
	3. Students set goals, monitor, and evaluate their reading progress	This standard is not measurable in a statewide assessment.		
	4. Students select, read, and respond to print and nonprint material for a variety of purposes	10		
	5. Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	10		

CRT Performance Level Descriptors

Advanced (289-300)
This level denotes superior performance.

Proficient (250-288)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2 CRT

Confidential

Reading **System Summary Report**

System:
Grade: 04
Spring 2008

III. Results for Subgroups of Students

Reporting Category	System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students										
Gender										
Male										
Female										
Ethnicity										
American Indian or Alaska Native										
Asian										
Hispanic										
Black or African American										
Native Hawaiian or Other Pacific Islander										
White										
Special Education										
Students with a 504 Plan										
Title I (optional)										
Tested with Standard Accommodation										
Tested with Non-Standard Accommodation										
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report								
Migrant										
Gifted/Talented										
LEP/ELL										
Former LEP Student										
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students								
Free/Reduced Lunch										
Significant Cognitive Disability										
Special Education Disability(ies):										
Autism										
Cognitive Delay										
Deaf-Blindness Impairment										
Deafness										
Emotional Disturbance										
Hearing Impairment										
Learning Disability										
Other Health Impairment										
Orthopedic Impairment										
Speech/Language										
Traumatic Brain Injury										
Visual Impairment										

Data not available for the 2008 report

MontCAS, Phase 2
CRT

System:
Grade: 04
Spring 2008

Science

System Summary Report

I. Distribution of Scores

Perf. Level	Scores	System			State		
		Number	% of Students	% of Students in Cat.	Number	% of Students	% of Students in Cat.
Advanced	297-300						
	293-296						
	289-292						
	285-288						
	281-284						
Proficient	275-280						
	269-274						
	262-268						
	256-261						
	250-255						
Nearing Proficiency	245-249						
	240-244						
	235-239						
	230-234						
	225-229						
Novice	220-224						
	215-219						
	210-214						
	205-209						
	200-204						

II. Subtest Results

Science		Possible Points	Average Points Earned	
			System	State
Total Points		61		
Standards	1. Scientific Investigations	14		
	2. Physical Science	14		
	3. Life Science	14		
	4. Earth and Space Science	14		
	5. Impact on Society	Sub scores are not reported for this standard		
	6. Historical Development	Sub scores are not reported for this standard		

CRT Performance Level Descriptors

Advanced (281-300)
This level denotes superior performance.

Proficient (250-280)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2 CRT

Confidential

Science **System Summary Report**

System:
Grade: 04
Spring 2008

III. Results for Subgroups of Students

Reporting Category	System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students										
Gender										
Male										
Female										
Ethnicity										
American Indian or Alaska Native										
Asian										
Hispanic										
Black or African American										
Native Hawaiian or Other Pacific Islander										
White										
Special Education										
Students with a 504 Plan										
Title I (optional)										
Tested with Standard Accommodation										
Tested with Non-Standard Accommodation										
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report								
Migrant										
Gifted/Talented										
LEP/ELL										
Former LEP Student										
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students								
Free/Reduced Lunch										
Significant Cognitive Disability										
Special Education Disability(ies):										
Autism										
Cognitive Delay										
Deaf-Blindness Impairment										
Deafness										
Emotional Disturbance										
Hearing Impairment										
Learning Disability										
Other Health Impairment										
Orthopedic Impairment										
Speech/Language										
Traumatic Brain Injury										
Visual Impairment										

Data not available for the 2008 report

* Less than ten (10) students were assessed

MontCAS, Phase 2
CRT

School:
System:
Grade: 04
Spring 2008

Mathematics

School Summary Report

I. Distribution of Scores

Perf. Level	Scores	School			System			State		
		N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.
Advanced	299-300									
	297-298									
	295-296									
	293-294									
	291-292									
Proficient	283-290									
	275-282									
	266-274									
	258-265									
	250-257									
Nearing Proficiency	245-249									
	240-244									
	235-239									
	230-234									
	225-229									
Novice	220-224									
	215-219									
	210-214									
	205-209									
	200-204									

II. Subtest Results

Mathematics		Possible Points	Average Points Earned		
			School	System	State
Total Points		66			
Standards	1. Problem Solving	This standard is assessed within the frameworks of standards 2-7.			
	2. Numbers and Operations	22			
	3. Algebra	8			
	4. Geometry	10			
	5. Measurement	10			
	6. Data Analysis, Statistics, and Probability	8			
	7. Patterns, Relations, and Functions	8			

CRT Performance Level Descriptors

Advanced (291-300)
This level denotes superior performance.

Proficient (250-290)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2 CRT

Confidential

Mathematics **School Summary Report**

School:
System:
Grade: 04
Spring 2008

III. Results for Subgroups of Students

Reporting Category	School					System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students															
Gender															
Male															
Female															
Ethnicity															
American Indian or Alaska Native															
Asian															
Hispanic															
Black or African American															
Native Hawaiian or Other Pacific Islander															
White															
Special Education															
Students with a 504 Plan															
Title I (optional)															
Tested with Standard Accommodation															
Tested with Non-Standard Accommodation															
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report													
Migrant															
Gifted/Talented															
LEP/ELL															
Former LEP Student															
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students													
Free/Reduced Lunch															
Significant Cognitive Disability	Data not available for the 2008 report														
Special Education Disability(ies):															
Autism															
Cognitive Delay															
Deaf-Blindness Impairment															
Deafness															
Emotional Disturbance															
Hearing Impairment															
Learning Disability															
Other Health Impairment															
Orthopedic Impairment															
Speech/Language															
Traumatic Brain Injury															
Visual Impairment															

* Less than ten (10) students were assessed

MontCAS, Phase 2
CRT

School:
System:
Grade: 04
Spring 2008

Reading

School Summary Report

I. Distribution of Scores

Perf. Level	Scores	School			System			State		
		N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.
Advanced	299-300									
	296-298									
	294-295									
	291-293									
	289-290									
Proficient	281-288									
	273-280									
	266-272									
	258-265									
	250-257									
Nearing Proficiency	245-249									
	240-244									
	235-239									
	230-234									
	225-229									
Novice	220-224									
	215-219									
	210-214									
	205-209									
	200-204									

II. Subtest Results

Reading		Possible Points	Average Points Earned		
			School	System	State
Total Points		60			
Standards	1. Students construct meaning as they comprehend, interpret, and respond to what they read	21			
	2. Students apply a range of skills and strategies to read	19			
	3. Students set goals, monitor, and evaluate their reading progress	This standard is not measurable in a statewide assessment.			
	4. Students select, read, and respond to print and nonprint material for a variety of purposes	10			
	5. Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	10			

CRT Performance Level Descriptors

Advanced (289-300)
This level denotes superior performance.

Proficient (250-288)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2 CRT

Confidential

Reading **School Summary Report**

School:
System:
Grade: 04
Spring 2008

III. Results for Subgroups of Students

Reporting Category	School					System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students															
Gender															
Male															
Female															
Ethnicity															
American Indian or Alaska Native															
Asian															
Hispanic															
Black or African American															
Native Hawaiian or Other Pacific Islander															
White															
Special Education															
Students with a 504 Plan															
Title I (optional)															
Tested with Standard Accommodation															
Tested with Non-Standard Accommodation															
Alternate Assessment		**If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report**													
Migrant															
Gifted/Talented															
LEP/ELL															
Former LEP Student															
LEP Student Enrolled for First Time in a U.S. School		**Performance levels are not reported for 1st year LEP students**													
Free/Reduced Lunch															
Significant Cognitive Disability	**Data not available for the 2008 report**														
Special Education Disability(ies):															
Autism															
Cognitive Delay															
Deaf-Blindness Impairment															
Deafness															
Emotional Disturbance															
Hearing Impairment															
Learning Disability															
Other Health Impairment															
Orthopedic Impairment															
Speech/Language															
Traumatic Brain Injury															
Visual Impairment															

* Less than ten (10) students were assessed

MontCAS, Phase 2
CRT

School:
System:
Grade: 04
Spring 2008

Science

School Summary Report

I. Distribution of Scores

Perf. Level	Scores	School			System			State		
		N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.
Advanced	297-300									
	293-296									
	289-292									
	285-288									
	281-284									
Proficient	275-280									
	269-274									
	262-268									
	256-261									
	250-255									
Nearing Proficiency	245-249									
	240-244									
	235-239									
	230-234									
	225-229									
Novice	220-224									
	215-219									
	210-214									
	205-209									
	200-204									

II. Subtest Results

Science		Possible Points	Average Points Earned		
			School	System	State
Total Points		61			
Standards	1. Scientific Investigations	14			
	2. Physical Science	14			
	3. Life Science	14			
	4. Earth and Space Science	14			
	5. Impact on Society	Sub scores are not reported for this standard			
	6. Historical Development	Sub scores are not reported for this standard			

CRT Performance Level Descriptors

Advanced (281-300)
This level denotes superior performance.

Proficient (250-280)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2 CRT

Confidential

Science **School Summary Report**

School:
System:
Grade: 04
Spring 2008

III. Results for Subgroups of Students

Reporting Category	School					System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students															
Gender															
Male															
Female															
Ethnicity															
American Indian or Alaska Native															
Asian															
Hispanic															
Black or African American															
Native Hawaiian or Other Pacific Islander															
White															
Special Education															
Students with a 504 Plan															
Title I (optional)															
Tested with Standard Accommodation															
Tested with Non-Standard Accommodation															
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report													
Migrant															
Gifted/Talented															
LEP/ELL															
Former LEP Student															
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students													
Free/Reduced Lunch															
Significant Cognitive Disability	Data not available for the 2008 report														
Special Education Disability(ies):															
Autism															
Cognitive Delay															
Deaf-Blindness Impairment															
Deafness															
Emotional Disturbance															
Hearing Impairment															
Learning Disability															
Other Health Impairment															
Orthopedic Impairment															
Speech/Language															
Traumatic Brain Injury															
Visual Impairment															

* Less than ten (10) students were assessed

C o n f i d e n t i a l

Roster and Item-Level Report
Mathematics

System:
School:
Grade: 4
Date: 9/19/2008 Page: 1

Last Name	First Name	Position	3	12	13	15	19	23	24	25	28	35	37	42	43	48	61	63	65	69	70	Scaled Score	Performance Level
		Standard	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2		
		Key	D	C	C	A	C				B	C	C	D	C		A	B	A	B	A		
		Points Possible	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1		
			B	A	+	B	+	1	1	4	+	+	+	B	+	1	+	+	+	+	+	300	A
			+	+	+	+	+	1	1	4	+	+	+	+	+	1	+	+	+	+	+	300	A
			+	+	+	D	+	0	1	1	+	+	D	+	D	0	B	+	B	+	D	276	P
			+	+	+	B	+	0	1	4	+	+	+	+	+	0	+	+	B	+	+	300	A
			+	+	+	+	A	1	1	1	+	+	D	+	B	0	C	A	B	+	+	255	P
			+	+	+	+	+	1	D	3	+	+	+	+	B	0	+	A	B	+	+	289	P
			A	+	D	D	B	0	0	2	D	+	A	B	B	0	+	+	B	C	+	219	N
			+	+	B	+	+	0	0	3	+	+	B	+	+	1	+	+	+	+	+	286	P
			+	+	+	+	+	1	1	3	+	+	+	+	+	1	+	+	+	C	+	300	A
			A	+	+	+	+	0	0	1	+	A	B	A	+	0	D	D	C	C	+	219	N
			C	+	B	C	+	0	0	2	C	+	B	A	+	0	+	C	C	+	C	202	N
			+	+	+	+	+	1	1	2	+	+	A	+	D	0	+	D	B	+	C	276	P
			+	+	+	C	A	0	1	1	+	A	B	+	D	0	C	B	+	+	C	240	NP
			+	+	+	+	+	1	1	3	+	+	+	+	+	1	+	+	+	+	+	299	A
			+	+	A	+	+	1	1	4	+	+	+	+	+	1	+	+	+	A	+	300	A
			+	+	+	+	D	0	0	3	A	+	D	+	B	0	+	+	B	C	D	255	P
			+	A	+	+	+	1	1	3	C	+	+	+	+	1	+	+	B	+	+	300	A
			+	+	+	+	+	0	1	3	+	A	+	+	+	1	+	+	+	D	B	280	P
			B	+	B	+	+	1	1	4	+	+	+	+	+	1	+	+	B	+	+	300	A
			+	+	+	+	+	0	1	3	+	+	+	B	+	0	+	+	+	+	D	293	A
			A	+	A	C	B	0	0	3	+	A	+	+	B	1	+	+	C	+	C	270	P
			C	+	+	B	A	0	0	2	D	B	+	A	B	0	+	D	+	+	+	216	N
			C	+	+	+	B	1	1	4	+	+	+	+	+	0	+	+	+	C	+	289	P
			+	+	+	+	+	0	D	4	+	+	+	+	+	1	+	+	+	+	D	300	A
			C	+	+	+	D	0	0	2	D	D	+	+	+	0	C	+	B	+	C	246	NP

Confidential

Roster and Item-Level Report
Reading

System:
School:
Grade: 4
Date: 9/19/2008 Page: 1

		Position	2	3	17	18	19	30	31	32	33	35	40	56	57	59	60	69	70	71	73	75	77		Scaled Score	Performance Level
		Standard	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
		Key	B	A	A	C	B	C	D	D	B	B	B	B	A	D	D	A	B	D	A	C	A			
		Points Possible	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
Last Name	First Name																									
			+	+	+	+	+	+	+	+	+	C	A	+	+	+	+	+	+	+	+	B	+	B		291	A
			+	+	C	A	C	A	+	+	+	C	+	+	+	B	A	+	+	+	+	+	+		281	P
			+	+	+	+	+	+	+	+	+	+	+	+	C	+	+	+	+	+	B	+	D		285	P
			+	+	D	+	+	B	+	+	+	A	+	+	+	+	B	+	+	+	C	+	+		295	A
			+	D	B	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	D	D		284	P
			A	+	D	+	D	+	+	+	+	C	C	+	+	+	B	+	C	+	+	B	+		263	P
			C	C	+	+	A	+	+	+	+	+	A	+	+	+	B	+	A	+	+	D	+		273	P
			+	C	D	+	+	+	+	A	+	A	A	+	+	+	+	+	C	+	+	+	+		284	P
			+	+	+	+	+	+	+	+	+	A	C	+	+	+	+	+	D	+	+	+	+		290	A
			+	+	+	+	+	+	+	A	+	C	A	+	+	A	A	C	A	B	B	+	+		247	NP
			+	+	B	+	D	+	B	+	C	+	+	D	D	B	C	+	+	+	B	D	B		235	NP
			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	C	B	C		288	P
			+	+	+	B	+	+	B	A	+	A	D	+	+	+	C	B	C	+	B	+	+		263	P
			+	+	+	+	+	+	+	A	+	D	+	+	+	+	+	D	C	+	B	+	+		285	P
			+	+	+	+	+	+	+	B	+	+	+	+	+	+	+	+	+	+	+	+	+		300	A
			+	C	D	B	+	+	C	+	A	C	C	+	+	+	C	B	C	C	B	+	C		244	NP
			+	+	+	U	+	U	+	A	+	L	A	+	+	+	A	U	A	+	B	+	U		273	P
			+	+	C	+	+	+	+	A	+	A	+	+	+	+	+	+	+	+	+	+	B		288	P
			+	+	+	+	+	+	+	A	+	+	+	+	+	+	+	+	+	+	+	+	+		300	A
			+	+	+	+	+	+	+	+	+	C	+	A	+	+	+	+	+	C	B	D	B		281	P
			+	D	+	+	+	+	+	+	+	A	+	+	+	+	+	B	+	+	+	+	+		284	P
			+	C	+	A	A	+	+	A	+	D	D	+	+	A	+	D	D	+	B	B	C		234	NP
			+	C	+	+	+	D	+	+	+	A	+	A	+	+	B	+	+	+	+	+	+		294	A
			+	+	+	+	+	+	+	A	+	A	D	+	+	+	+	+	+	+	B	+	B		290	A
			+	C	D	+	+	D	+	A	C	C	+	+	+	A	B	C	C	+	B	D	+		244	NP

Confidential

Roster and Item-Level Report
Science

System:
School:
Grade: 4
Date: 9/19/2008 Page: 1

		Position	5	9	37	38	43	58	71	72	78	80	81						Scaled Score	Performance Level
		Standard	1	1	1	1	1	1	1	1	1	1	1							
		Key	A	D	A	B	B	D	A	D	A	C								
		Points Possible	1	1	1	1	1	1	1	1	1	1	4							
Last Name	First Name																			
			+	+	+	+	+	+	+	+	+	+	1						283	A
			+	+	+	+	+	+	C	+	+	+	3						273	P
			+	B	+	+	D	+	+	+	+	+	2						273	P
			+	+	D	+	+	+	+	+	+	+	2						249	NP
			+	+	D	D	+	+	+	+	+	+	2						262	P
			+	+	D	+	+	+	+	+	+	+	2						249	NP
			+	C	B	+	+						B						217	N
			+	C	+	A	D	+	+	B	+	+	1						249	NP
			+	B	D	C	+	+	+	+	+	+	1						259	P
			+	+	+	A	A	C	B	+	+	+	0						243	NP
			+	+	+	D	+	+	+	C	D	+	1						232	NP
			+	+	D	+	C	+	+	+	+	+	1						267	P
			+	+	D	+	+	C	B	B	C	A	1						219	N
			+	+	D	+	+	+	+	+	+	+	0						273	P
			+	B	+	+	+	+	+	+	+	+	3						296	A
			+	+	B	C	+	+	D	+	C	+	1						238	NP
			+	+	D	+	+	+	+	+	+	+	1						264	P
			+	A	+	C	+	+	+	+	+	+	2						256	P
			+	+	B	+	+	+	+	+	+	+	3						291	A
			+	A	+	+	+	+	+	+	+	+	3						273	P
			C	+	+	+	+	+	B	A	C	+	1						252	P
			+	C	+	C	D	+	+	A	D	B	1						240	NP
			+	+	D	+	+	+	+	+	B	+	1						256	P
			+	+	B	+	+	+	+	+	+	+	3						267	P
			+	+	+	+	+	+	+	+	B	+	0						259	P

CRT Performance Level Descriptors

The Performance Level Descriptors below describe students' knowledge, skills, and abilities in a content area. These descriptions provide a picture or profile of student achievement at the four performance levels: **Advanced**, **Proficient**, **Nearing Proficiency**, and **Novice**. Grade and content performance level descriptors may be found on OPI's web site at http://www.opi.mt.gov/assessment/index.html

Advanced
This level denotes superior performance.

Proficient
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

	Score Ranges		
	Reading	**Math**	**Science**
Advanced	(289-300)	(291-300)	(281-300)
Proficient	(250-288)	(250-290)	(250-280)
Nearing Proficiency	(225-249)	(225-249)	(225-249)
Novice	(200-224)	(200-224)	(200-224)

Reading Standards

1. Students construct meaning as they comprehend, interpret, and respond to what they read.
2. Students apply a range of skills and strategies to read.
3. Students set goals, monitor, and evaluate their reading progress.
4. Students select, read, and respond to print and nonprint material for a variety of purposes.
5. Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences.

Mathematics Standards

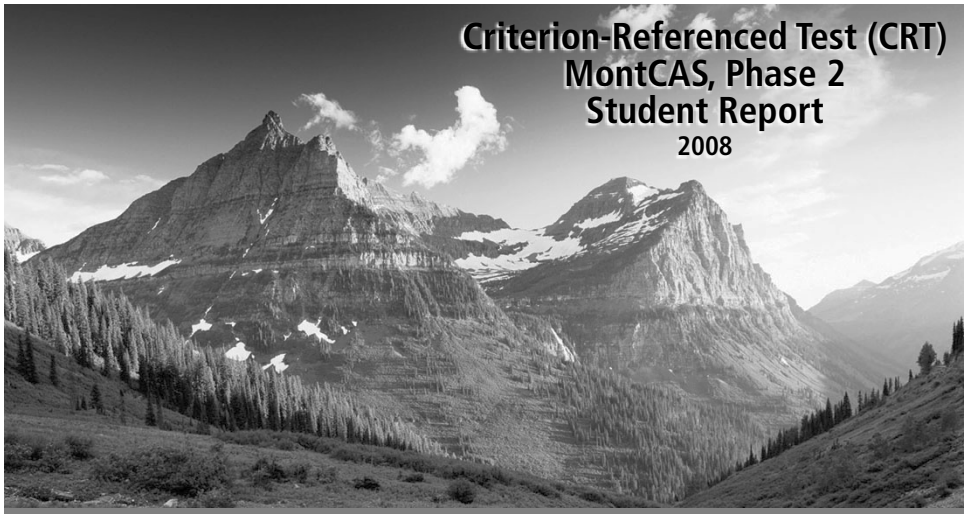
1. Problem Solving
2. Numbers and Operations
3. Algebra
4. Geometry
5. Measurement
6. Data Analysis, Statistics, and Probability
7. Patterns, Relations, and Functions

Science Standards

1. Scientific Investigations
2. Physical Science
3. Life Science
4. Earth/Space Science
5. Impact on Society
6. Historical Development

OPI Contact
Judy Snow, State Assessment Director
406-444-3656
jsnow@mt.gov

For more information regarding student assessments in Montana, check out the Office of Public Instruction's Parents Page at http://www.opi.mt.gov/parents.



Criterion-Referenced Test (CRT)
MontCAS, Phase 2
Student Report
2008

Student Name:
School:
System:
Grade: 04

Dear Parents/Guardians:

This report contains the results of the Spring 2008 Montana Comprehensive Assessment System (MontCAS) Criterion-Referenced Test (CRT) that your child took in March. The CRT provides schools with information to evaluate and improve curriculum and instruction to help all students meet Montana's content standards. This report provides important information about your child's performance on the assessment along with state results.

The CRT contains multiple-choice, short-answer questions, and constructed responses. The test measures a student's knowledge of subject matter identified in the Montana State Standards for Reading, Mathematics, and Science. Science is assessed in grades 4, 8, and 10 only.

It is important to remember that the CRT is just one measure of your child's academic progress. Your local school staff can provide further information about your child's performance in school. The CRT, which is required by the No Child Left Behind Act, is part of an ongoing statewide educational improvement process. Working together, we can ensure that Montana's children continue to receive a high-quality education.

Sincerely,

Linda McCulloch
Montana Superintendent of Public Instruction

Montana Office of Public Instruction
PO Box 202501
Helena, Montana 59620-2501
http://www.opi.mt.gov

How did do on the CRT?

Scaled Scores on the CRT

The criterion-referenced test (CRT) is designed to measure student performance against the learning goals described in the Montana Content Standards (http://www.opi.state.mt.us/standards/index.html). Consistent with this purpose, results on the CRT are reported according to performance levels that describe student performance in relation to the established state standards. There are four performance levels: **Advanced**, **Proficient**, **Nearing Proficiency**, and **Novice**. Your child's performance levels in reading, mathematics, and science* are based on a total scaled score in each content area. Scaled scores in each content area range from 200 to 300. Your child's performance levels, based on the scaled scores, are shown in the bar graphs below.

Subject Area	Performance Level	Scaled Score	Display of Score and Probable Range of Scores
Reading	Nearing Proficiency	226	Novice / Nearing Proficiency / Proficient / Advanced; 200 225 250 289 300; SCALED SCORE
Mathematics	Nearing Proficiency	240	Novice / Nearing Proficiency / Proficient / Advanced; 200 225 250 291 300; SCALED SCORE
Science*	Nearing Proficiency	249	Novice / Nearing Proficiency / Proficient / Advanced; 200 225 250 281 300; SCALED SCORE

Scores on Montana Content Standards

In addition to performance levels, CRT results are reported for Montana Content Standards in Reading, Mathematics, and Science. Unlike scaled scores which provide a total performance level score, Montana Content Standard Scores provide more specific information about your child's achievement on the CRT. The charts below show your child's performance compared to the overall state performance in each area of study within subject areas (Montana Content Standards for Reading, Math, and Science). These results can be used to show your child's relative strengths or weaknesses.

This Student's Performance Levels Relative to Student Achievement for State

	Reading		Mathematics		Science*	
	Student	State	Student	State	Student	State
Advanced		34		27		14
Proficient		45		40		48
Nearing Proficiency	✔	16	✔	19	✔	30
Novice		5		14		7

This Student's Performance in Content Area Standards

Reading	Total Possible Points	Student % of Points Earned	Points Earned: Average State %
Standard 1	21	48	68
Standard 2	19	21	62
Standard 3	This standard is not measurable in a statewide assessment.		
Standard 4	10	40	53
Standard 5	10	10	64

Science*	Total Possible Points	Student % of Points Earned	Points Earned: Average State %
Standard 1	14	64	67
Standard 2	14	64	70
Standard 3	14	79	76
Standard 4	14	50	65
Standard 5	Sub scores are not reported for this standard.		
Standard 6	Sub scores are not reported for this standard.		

Mathematics	Total Possible Points	Student % of Points Earned	Points Earned: Average State %
Standard 1	This standard is assessed within the frameworks of standards 2-7.		
Standard 2	22	45	59
Standard 3	8	50	62
Standard 4	10	50	62
Standard 5	10	40	63
Standard 6	8	63	77
Standard 7	8	63	58

The standards for each content area can be found on the back of this report.
*Science is assessed at grades 4, 8, and 10 only.

Contact your student's school for more information about the following symbols:
† Student did not complete the assessment. § Student participated with a non-standard accommodation. **Student did not participate.

APPENDIX F—REPORTING DECISION RULES

Analysis and Reporting Decision Rules
Montana Comprehensive Assessment System (MontCAS) CRT and CRT-Alternate (Final)
Spring 07-08 Administration

This document details rules for analysis and reporting. The final student level data set used for analysis and reporting is described in the "Data Processing Specifications." This document is considered a draft until the Montana Office of Public Instruction (OPI) signs off. If there are rules that need to be added or modified after said sign-off, OPI sign off will be obtained for each rule. Details of these additions and modifications will be in the Addendum section.

I. General Information

A. Tests Administered

Grade	Subject	Items included in Raw Score		IABS Reporting Categories (Standards) (Not Applicable for CRT-Alternate)
		CRT	CRT-Alt	
03	Reading Math	Common	All	Cat2
04	Reading Math	Common	All	Cat2
	Science	Common	All	Cat3
05	Reading Math	Common	All	Cat2
06	Reading Math	Common	All	Cat2
07	Reading Math	Common	All	Cat2
08	Reading Math	Common	All	Cat2
	Science	Common	All	Cat3
10	Reading Math	Common	All	Cat2
	Science	Common	All	Cat3

B. Reports Produced

1. Student Labels
2. Student Report
3. Roster & Item Level Report (online system)
   - by grade, subject and class/group
4. Summary Report
   Consists of sections:
   I. Distribution of Scores
   II. Subtest Results
   III. Results for Subgroups of Students
   - by grade, subject and school

- by grade, subject and system
- by grade, subject (state level)

C. Files Produced (excel file format)
   1. One state file for each grade
      a. Consists of student level results
      b. Alternately assessed students are in separate files by grade.
   2. Naming convention
      a. CRT Reading and Math- StudentdatafileReaMat[2 digit grade].xls
      b. CRT Science- StudentdatafileSci[2 digit grade].xls
      c. CRT-Alternate- altStudentdatafileReaMat[2 digit grade].xls
      d. CRT-Alternate- altStudentdatafileSci[2 digit grade].xls

D. School Type

Schtype	**Source**	**Description**	**Included in Aggregations**		
			School	**System**	**State**
"Pras"	Data file provided by state	Private Accredited School. They are their own system	Yes. Same information for school & system but both sets of reports produced	Yes. Same information for school & system but both sets of reports produced	No
"Prnas"	Scanned data	Private non-accredited school. They are their own system	Yes. Same information for school & system but both sets of reports produced	Yes. Same information for school & system but both sets of reports produced	No
"SNE"	Scanned data	Student not enrolled	No.	No.	No.
"Oth"	Data file provided by state/Scanned data	non-private school	Yes	Yes	Yes

E. Other Information
   1. CRT Tests are constructed with a combination of common and embedded matrix field test items.
   2. The CRT-Alternate consists of a set of performance tasks. At grades 3, 5, 6, and 7 the tasks are grouped into five (5) sets of five (5) tasklets for each subject. At grades 4, 8 and 10 (Reading and Math) the tasks are not grouped. At grades 4, 8 and 10 science is grouped into 5 tasklets. The number of activities in each tasklet varies.

II. Student Participation/Exclusions

A. Test Attempt Rules

1. A valid response to a multiple choice item is A, B, C, or D. An asterisk (multiple marks) is not considered a valid response.
2. Incomplete (CRT): The student has fewer than two (2) but at least one (1) valid responses to common multiple choice items.
3. Incomplete (CRT-Alternate): The student responded to fewer than three (3) items.
4. The student is classified as Did Not Participate (DNP) in CRT if the student does not have any valid responses for that subject in either CRT or CRT-Alternate.

B. Not Tested Reasons

N/A

C. Student Participation Status

1. The following students are excluded from all aggregations.
    a. Foreign Exchange Students (FXS).
    b. Homeschooled students (schtype='SNE').
    c. Part-time students (PSNE).
    d. DNP (for that subject)
    e. First year LEP
    f. Student tested with Non-Standard Accommodations (NSA for that subject)

2. If any of the non-standard accommodations are bubbled the student is considered tested with non-standard accommodations (NSA) in that subject.
3. If the student has not been in that school for the entire academic year the student is excluded from school level aggregations (NSAY).
4. If the student has not been in that system for the entire academic year the student is excluded from system and school level aggregations (NDAY).
5. If the student took the alternate assessment the student is not counted as participating in the general assessment. Alternate Assessment students receive their results on an Alternate Assessment Student Report. They are reported according to participation rules stated in this document.
6. (CRT-Alternate) If the teacher halted the administration of the assessment after the student scored zero (0) for three (3) consecutive items (within tasklets for grades 3, 5, 6, and 7 and science (grades 4, 8 and 10)) the student is classified as Halted in that subject. Scores received after three (3) consecutive zeroes are blanked out and are not counted toward the student's score. For grades 3,5,6,7 and science if the student was halted within a tasklet then the rest of the items within the tasklet are blanked out and do not count toward the student's score. If the other tasklets are complete then those items will be counted toward the student's score.

D. Student Participation Summary

Participation Status	Part. Flag	Raw score	Scaled Score	Perf. level	Included on Roster	Included in aggregations		
						Sch	Sys	Sta
FXS	E	Yes	Yes	Yes	No	No	No	No
SNE	E	Yes	Yes	Yes	No	No	No	No
PSNE	E	Yes	Yes	Yes	No	No	No	No
NSA(by subject)	A	Yes	Yes	Yes	Yes	No	No	No
First year LEP	A	Yes	See Report Specific Rules	See Report Specific Rules	Yes	Only in count of First year LEP		
NSAY only	B	Yes	Yes	Yes	Yes	No	Yes	Yes
NDAY	C	Yes	Yes	Yes	Yes	No	No	Yes
ALT*	A	Yes	Yes	Yes	Yes	See footnote below		
Incomplete	A	Yes	Yes	Yes	Yes	No	No	No
DNP(Non-Participants)	F	Yes	Yes	Yes	Yes	No	No	No
Halted(CRT-Alt only by subject)	D	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tested	Z	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* Alternate assessment students are included only in the count of alternate assessment students in general assessment reports. They are included in summary data only for alternate assessment reports (according to participation rules).

III. Calculations

A. Raw Scores
   1. (CRT) Raw scores are calculated using the scores on common multiple choice and open response items.
      (CRT-Alternate) Raw score is the sum of the individual item scores.
   2. Percentages and averages are reported to the nearest whole number.
   3. The number of included students (N) in a subject is the number of students in the school/system/state minus FXS minus PRAS minus PRNAS minus PSNE minus SNE minus First year LEP minus Incomplete minus NSA minus DNP.
   4. School/system reports are produced regardless of N-size.

B. Scaling
   Scaling is done using constants from psychometrics and the student's raw score.

C. Performance levels are assigned based on the student's earned raw score.

D. The classcode is created using the following steps:
   1. The following students are not included when creating the class codes.
      - SNE
      - ALT(CRT-only)

- FXS
- PSNE

2. The dataset (by grade) is sorted by schcode and class/group name
3. The records are then numbered consecutively starting at 1. This number is then padded with zeros (in front) to create a 3 digit number.

E. Performance Level coding:

Numeric Performance Level	Performance level Name	Abbreviation
1(lowest)	Novice	N
2	Nearing Proficient	NP
3	Proficient	P
4(highest)	Advanced	A

IV. Report Specific Rules

A. Student Label

1. If a student is First year LEP and incomplete in Reading, the Reading performance level is 'LEP'. The reading scaled score is blank.
2. If a student is First year LEP, the math and science performance levels are the name of the earned performance level and the scaled scores are the student's earned score.
3. If the student is not first year LEP, the performance level name corresponding to the student's earned score is displayed.
4. If the student is First year LEP but is not incomplete in Reading then the student receives his earned scaled score and performance level.
5. If the student is DNP the student receives a student label. The student receives scaled score =200 and performance level=Novice.

B. Student Report

1. If a student is First year LEP and incomplete in Reading the Reading performance level is 'LEP' and the scaled score is blank.

2. If the student is First year LEP but is not incomplete in Reading then the student receives his earned scaled score and performance level.
3. If a student is First year LEP, the math and science performance levels are the name of the earned performance level and the scaled score is the student's earned score.
4. If the student is not first year LEP, the performance level name corresponding to the student's earned score is displayed.
5. If the student is incomplete the student receives the scores with a footnote (†) "Student did not complete the assessment."
6. If the student is NSA the student will receive his scores with the footnote (§) "Student took non-standard accommodation."
7. There is no last name or first name for the student, the name displayed is "Name Not Provided".

8. Alt students who are halted receive their scores and performance level and a footnote(§)

a. Grades 4,8,10 Reading and Math "Teacher halted the administration of the assessment after the student scored a 0 for three consecutive items on different test administrations"
b. Grades 3,5,6,7 and Science "Teacher halted the administration of one or more of the five test activities after the student scored a 0 for three consecutive items within an activity on two different test administrations. Any completed test activities have been scored and are reflected in the student's scaled score."

9. If the student is DNP the student receives a Student Report. The student receives scaled score =200 and performance level =Novice. The standards will not be reported. The student receives a footnote (**) "Student did not participate."

C. Roster & Item Level Report
1. If a student is First year LEP and the student is not incomplete in Reading:
   a. The math (and science) performance level is the abbreviation of the earned performance level and the scaled score is the student's earned score.
   b. The reading performance level is the abbreviation of the earned performance level and the scaled score is the student's earned score.
   c. The student is excluded from Reading, Math and Science aggregations.
2. If the student is First year LEP and incomplete in Reading
   a. The student's Reading, Math (and Science) performance levels are 'LEP'.
   b. The student's math (and science) scaled score is the student's earned scaled score and the reading scaled score is blank.
   c. The student's responses for all subjects are displayed.
   d. The student is excluded from Math, Reading (and Science) aggregations.
3. If the student is not first year LEP, the performance level abbreviation corresponding to the student's earned score is displayed.
4. If the student is incomplete the student receives the scores with a footnote (†) "Student did not complete the assessment."
5. If the student is NSA the student will receive his scores with the footnote (§) "Student took non-standard accommodation."
6. There is no last name or first name for the student, the name displayed is "Name Not Provided".
7. If class/group information is missing the roster is done at the school level.
8. Alternate Assessment students are reported only on their class/group/school's alternate *Roster & Item Level Report.*
9. If the student is a Non-Participant the student is listed on the *Roster & Items level Report.* All responses and scores will be blank. The scaled score =200 and performance level=N. The student will receive the footnote "Student did not participate in assessment."

D. School Summary
1. Section III (Results for Subgroups of Students)
   a. Performance level results for subgroups with N less than 10 are suppressed. N is always reported. Footnote * 'Less than 10 students were assessed.'
   b. Count of students who are considered NSA for that subject excluding those students who are incomplete, nsay (at school level), nday (at school and system level) or FXS or SNE or PSNE or First year LEP or alt (general assessment report).

c. Count of students who are alt excludes those students who are nsay (at school level), nday (at school or system level) or incomplete or FXS or SNE or PSNE or NSA or First year LEP.

d. Count of First year LEP students excludes those students who are nsay (at school level), nday (at school or system level) or incomplete or FXS or SNE or PSNE or NSA or alt (general assessment).

V. Data File Rules(Excel format**)**

1. The following students are not included in the state file
   a. Alternate Assessment students (in CRT)
   b. Homeschooled students(SNE)
   c. Part-Time students (PSNE)
2. If the student receives a performance level 'LEP' on the student report in Reading, the student receives LEP for the Reading performance level in the state files.
3. Alt students who are halted are marked '1' in the halted field for that subject.

Addenda

A. The following rules pertain to the Reporting Online Tool only. This section replaces Section IV.C *Roster & Item Level Report*.

1. Students who test with Non-Standard Accommodations (NSA) are included in school, system and state level aggregations.
2. Students who are NSAY are included in school, system and state level aggregations.
3. Students who are NDAY are included in school, system and state level aggregations.
4. Students who are DNP in a subject are reported with scaled score=200 and performance level ='DNP' on the interactive roster.
5. Students who are Incomplete in a subject are reported with their earned scaled score and performance level='INC' on the interactive roster.
6. Students who are first year LEP and who complete the Reading test are reported with their earned scaled score and performance level and are included in school, system and state level aggregations for all subjects unless otherwise excluded based on completeness in math or science.
7. Students who are first year LEP and who do not complete the Reading test are reported with their earned scaled score and performance level='LEP' for all subjects. These students are excluded from school, system and state level aggregations.
8. Students who participated in Alternate assessment are listed on the rosters. Their scaled score is blank and the performance level='ALT'. These students are not included in aggregations.

B. The following hierarchy serves to clarify the assignments of the partstatuses:

F (Student attempted no common items and is not alt)
E (FXS, SNE, PSNE)
A (NSA, LEPFirst, ALT, INC)
C (NDAY)

B (NSAY)
Z (completed CRT and none of the above conditions apply)

C. The data files (state, system, school) as applicable are in csv (comma delimited) format with csv extension.

D. The following students are not included in System level files:
- SNE
- PSNE
- ALT (excluded from CRT system files)
- FXS

E. LEPFirst students who have less than 2 (including zero (0)) valid responses in reading receive performance level='LEP' on the roster for all subjects. If an LEPFirst student attempts 2 or more items in Reading they receive their earned performance level for Reading. They receive their Math (and Science) performance level based on completeness of these tests.

F. Alt roster rules are as stated in Section IV.C *Roster & Item Level Report.* The Roster addendum above pertains to CRT only.

G. The scaled scores for students with performance level='LEP' on the interactive roster are reported.

H. On the interactive site the items on the roster are reported using the test positions.

I. The format for the system level files for the static side of the online system is excel.

APPENDIX G—SUBGROUP RELIABILITIES

Table G-1. 2007-08 MT CRT: Reliabilities of Subgroups by Grade and Subject.

Grade	Subject	Subgroup	N	(α)
3	Math	White	8518	0.90
		Native Hawaiian or Pacific Islander	19	0.85
		Hispanic or Latino	288	0.89
		Black or African American	131	0.92
		Asian	80	0.88
		American Indian or Alaskan Native	1291	0.90
		LEP	488	0.90
		IEP	1126	0.92
		Low SES	4250	0.90
	Reading	White	8502	0.89
		Native Hawaiian or Pacific Islander	19	0.79
		Hispanic or Latino	285	0.89
		Black or African American	132	0.90
		Asian	80	0.84
		American Indian or Alaskan Native	1294	0.89
		LEP	491	0.88
		IEP	1109	0.91
		Low SES	4246	0.90
4	Math	White	8638	0.90
		Native Hawaiian or Pacific Islander	20	0.84
		Hispanic or Latino	277	0.89
		Black or African American	119	0.89
		Asian	92	0.89
		American Indian or Alaskan Native	1201	0.90
		LEP	417	0.89
		IEP	1169	0.91
		Low SES	4142	0.90
	Reading	White	8612	0.89
		Native Hawaiian or Pacific Islander	20	0.90
		Hispanic or Latino	274	0.89
		Black or African American	118	0.88
		Asian	92	0.89
		American Indian or Alaskan Native	1205	0.89
		LEP	418	0.85
		IEP	1145	0.90
		Low SES	4129	0.90
	Science	White	8633	0.84
		Native Hawaiian or Pacific Islander	20	0.80
		Hispanic or Latino	277	0.83
		Black or African American	118	0.83
		Asian	92	0.85
		American Indian or Alaskan Native	1205	0.85
		LEP	419	0.81
		IEP	1169	0.86
		Low SES	4141	0.85

Table F-*G*. 2007-08 MT CRT: Reliabilities of Subgroups by Grade and Subject

Grade	Subject	Subgroup	N	(α)
5	Math	White	8584	0.89
		Native Hawaiian or Pacific Islander	30	0.86
		Hispanic or Latino	296	0.90
		Black or African American	119	0.89
		Asian	73	0.89
		American Indian or Alaskan Native	1195	0.89
		LEP	388	0.87
		IEP	1144	0.89
		Low SES	4009	0.90
	Reading	White	8568	0.91
		Native Hawaiian or Pacific Islander	29	0.87
		Hispanic or Latino	297	0.92
		Black or African American	118	0.90
		Asian	69	0.92
		American Indian or Alaskan Native	1191	0.91
		LEP	381	0.90
		IEP	1120	0.91
		Low SES	3989	0.92
6	Math	White	8890	0.90
		Native Hawaiian or Pacific Islander	24	0.90
		Hispanic or Latino	294	0.91
		Black or African American	96	0.89
		Asian	124	0.91
		American Indian or Alaskan Native	1165	0.89
		LEP	404	0.87
		IEP	1153	0.87
		Low SES	4011	0.90
	Reading	White	8903	0.89
		Native Hawaiian or Pacific Islander	24	0.92
		Hispanic or Latino	292	0.91
		Black or African American	96	0.89
		Asian	123	0.89
		American Indian or Alaskan Native	1166	0.90
		LEP	403	0.87
		IEP	1161	0.89
		Low SES	4019	0.90
7	Math	White	9005	0.90
		Native Hawaiian or Pacific Islander	28	0.91
		Hispanic or Latino	259	0.89
		Black or African American	92	0.87
		Asian	105	0.91
		American Indian or Alaskan Native	1143	0.88
		LEP	416	0.85
		IEP	1145	0.85
		Low SES	3774	0.88
	Reading	White	9014	0.91
		Native Hawaiian or Pacific Islander	28	0.85
		Hispanic or Latino	259	0.92
		Black or African American	92	0.86
		Asian	101	0.91
		American Indian or Alaskan Native	1143	0.92
		LEP	412	0.88
		IEP	1154	0.90
		Low SES	3779	0.92

Table G-1. 2007-08 MT CRT: Reliabilities of Subgroups by Grade and Subject

Grade	Subject	Subgroup	N	(α)
8	Math	White	9275	0.91
		Native Hawaiian or Pacific Islander	32	0.88
		Hispanic or Latino	265	0.90
		Black or African American	124	0.89
		Asian	109	0.92
		American Indian or Alaskan Native	1131	0.89
		LEP	437	0.85
		IEP	1242	0.85
		Low SES	3926	0.90
	Reading	White	9291	0.90
		Native Hawaiian or Pacific Islander	32	0.84
		Hispanic or Latino	266	0.91
		Black or African American	125	0.88
		Asian	110	0.87
		American Indian or Alaskan Native	1140	0.91
		LEP	440	0.89
		IEP	1256	0.89
		Low SES	3940	0.91
	Science	White	9283	0.86
		Native Hawaiian or Pacific Islander	30	0.83
		Hispanic or Latino	268	0.86
		Black or African American	125	0.85
		Asian	110	0.84
		American Indian or Alaskan Native	1135	0.86
		LEP	437	0.81
		IEP	1265	0.85
		Low SES	3939	0.87
10	Math	White	9573	0.90
		Native Hawaiian or Pacific Islander	26	0.86
		Hispanic or Latino	245	0.90
		Black or African American	74	0.86
		Asian	101	0.91
		American Indian or Alaskan Native	1060	0.85
		LEP	331	0.69
		IEP	1104	0.76
		Low SES	3060	0.87
	Reading	White	9558	0.89
		Native Hawaiian or Pacific Islander	26	0.91
		Hispanic or Latino	246	0.90
		Black or African American	74	0.88
		Asian	101	0.89
		American Indian or Alaskan Native	1065	0.89
		LEP	333	0.83
		IEP	1089	0.87
		Low SES	3057	0.89
	Science	White	9560	0.88
		Native Hawaiian or Pacific Islander	26	0.92
		Hispanic or Latino	245	0.89
		Black or African American	74	0.87
		Asian	101	0.90
		American Indian or Alaskan Native	1061	0.85
		LEP	331	0.75
		IEP	1121	0.82
		Low SES	3061	0.87

[1]Only subgroups with sample size ≥10 reported

APPENDIX H—ACCOMODATIONS

Accommodations Selection Guidance

Type of Accommodation	ELL Students		Students with IEPs	
Scheduling Accommodations	**Direct**	**Indirect**	**Primary**	**Support**
1. Change in Administration Time: Test is administered at a time of day or a day of the week based on student needs.			X	x
2. Session Duration: Test is administered in appropriate blocks of time for individual student needs, followed by rest breaks.			X	
3. Extended Time: Time is extended beyond the regular test administration allotments until, in the administrator's judgment, the student could no longer sustain the activity.	X	x	X	x
Setting Accommodations	**Direct**	**Indirect**	**Primary**	**Support**
4. Individual Administration: Test was administered in a one to one situation.		x	X	x
5. Small Group Administration: Test was administered to a small group of students.		x	X	x
6. Reduce Distractors: Student is seated at a carrel or other physical arrangement that reduces visual distraction.			X	x
7. Alternative Setting: Test is administered to the student in a different setting.		x	X	x
8. Change in Personnel: Test is administered by other personnel known to the student (e.g., LEP, Title I, special education teacher).	X		X	x
9. Home Setting: Test is administered to the student by school personnel in their home.			X	
10. Front Row Seating: A student is seated in front of the classroom when taking the test.	X		X	x
11. Teacher Presence: A teacher faces the student during test administration.			X	x

Type of Accommodation	ELL Students		Students with IEPs	
Equipment Accommodations	**Direct**	**Indirect**	**Primary**	**Support**
12. Magnification: Student used equipment to magnify test materials.			X	
13. Noise Buffers: Student wears equipment to reduce environmental noises.			X	
14. Template: Student uses a template.			X	
15. Amplification: Student uses amplification equipment (e.g., hearing aid or auditory trainer) while taking test.			X	
16. Writing Tools: Student uses a typewriter or word processor (without activating spellchecker).			X	
17. Voice Activation: Student speaks response into computer equipped with voice activation software.			X	
18. Bilingual Dictionary: Student uses a bilingual dictionary (Note: Bilingual dictionary could include a simplified English dictionary or glossary, subject area vocabulary list).	X			
Recording Accommodations	**Direct**	**Indirect**	**Primary**	**Support**
19. Dictation: The student dictates answers to a test administrator who records them in the Test Booklet.			X	
20. Writing Tools: The student marks or writes answers with the assistance of a technology device or special equipment. The students' answers are transferred by the test administrator to the Test Booklet.			X	
21. Assistive Technology: Another form of assistive technology routinely used by the student (that does not change the intent or content of the test) was used by the student.			X	
Modality Accommodations	**Direct**	**Indirect**	**Primary**	**Support**

Type of Accommodation	ELL Students		Students with IEPs	
22. Oral Presentation: Tests were read to the student by the test administrator (with the exception of reading passages). Note: Readers must read test items/questions to the student word-for-word exactly as written. Readers may not clarify, elaborate, or provide assistance to the student regarding the meaning of words, intent of test questions, or responses to test items/questions.	X		X	
23. Test Interpretation: Tests, including directions, were interpreted for students who are deaf or hearing-impaired (with the exception of interpreting the reading test).			X	
24. Test Directions with Verification: An administrator gave test directions with verification (by using a highlighter) that the student understood them.	X		X	
25. Test Directions Support: An administrator assisted students in understanding test directions, including giving directions in native language.	X		X	
26. Sheltered English: Test was read to an LEP student in "sheltered English" (with the exception of reading the reading test).	X			
27. Braille: A braille version of the test was used by the student.			X	
28. Large Print: A large print version of the test was used by the student.			X	
29. Other: With verification from OPI in advance of the testing window, some other approved accommodation was used by a student.	X		X	